Predicting Protein Molecular Function

Barbara Elizabeth Engelhardt

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2007-171
http://www.eecs.berkeley.edu/Pubs/TechRpts/2007/EECS-2007-171.html

December 20, 2007

Predicting Protein Molecular Function

by

Barbara Elizabeth Engelhardt

B.S. (Stanford University) 1999
M.S. (Stanford University) 1999

A dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Computer Science

and the Designated Emphases in

Computational and Genomic Biology
and
Communication, Computation, and Statistics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Michael I. Jordan, Chair
Professor Richard M. Karp
Professor Steven E. Brenner

Fall 2007

The dissertation of Barbara Elizabeth Engelhardt is approved:

Professor Michael I. Jordan, Chair — Date

Professor Richard M. Karp — Date

Professor Steven E. Brenner — Date

University of California, Berkeley

Fall 2007

Predicting Protein Molecular Function

by

Barbara Elizabeth Engelhardt

Abstract

Predicting Protein Molecular Function

by

Barbara Elizabeth Engelhardt

Doctor of Philosophy in Computer Science

and the Designated Emphases in

Computational and Genomic Biology

and Communication, Computation, and Statistics

University of California, Berkeley

Professor Michael I. Jordan, Chair

The number of known nucleotide sequences encoding proteins is growing at an extraordinarily fast rate due to technologies developed in the last decade that enable rapid sequence acquisition. Such rapid acquisition is a prelude to understanding the molecular function and tertiary structure of these protein sequences, and from there to an understanding of the role these proteins play in a particular organism. The experimental technologies that enable us to understand molecular function have not progressed as fast as those for sequencing. One role of computational biology is to accurately predict protein molecular function based on the protein's sequence alone.

Phylogenomics is a field of study that approaches the problem of protein molecular function prediction from an evolutionary perspective. In particular, a phylogenomic analysis transfers existing (but sparse) molecular function annotations to a query protein based on a reconciled phylogeny, which explicitly represents the evolutionary relationships of a set of related proteins. In my dissertation, I formalize the phylogenomics methodology as a statistical graphical model of molecular function evolution. Within this framework, we can predict protein molecular function from protein sequence alone. Molecular function evolution is represented as a simple continuous

time Markov chain, and the random variables at each node in the tree are a subset of functional terms from the Gene Ontology. The model is encapsulated in a framework called SIFTER (Statistical Inference of Function Through Evolutionary Relationships).

SIFTER has performed well on a number of diverse protein families, as compared to standard annotation transfer methods and other phylogenomics-based approaches. SIFTER has been applied to the complete genomes of 46 fungal species, and is able to make molecular function predictions for a large percentage of the predicted proteins in these genomes. Moreover, through these predictions we can explore some genomic comparisons for fungi. Motivated by the high cost of characterization experiments, active learning techniques have also been applied to SIFTER's protein function predictions, with good results.

Professor Michael I. Jordan, Chair Date

Acknowledgements

My studies at Berkeley were flanked by the two worst phone calls of my life, both from my mother. In the first, my mother called to tell me she was watching one of the twin towers burn after a plane had hit it from her office building, and she said she was scared; I told her I'm sure everything was going to be fine and that life would continue normally, but of course I was wrong. The second, on March 23, 2007, she called to tell me that my father had been diagnosed with pancreatic cancer, but that everything was going to be fine and that life was would continue normally. In this instance, she was wrong. During the same period, I also was married and had my first child, two of the greatest events of my life. I mention these events to show that my growth and development in graduate school were not confined to what was happening within the walls of Berkeley, and I've found that my academic pursuits do not take place in ivory towers. For that I am grateful.

First and foremost, I would like to thank my advisor, Michael Jordan, for being the very best teacher and mentor I can imagine, in part by acting with the highest integrity in every aspect of his advisor's role. I would like to thank Steven Brenner for introducing me to the exciting field of bioinformatics and being a supportive collaborator. I would also like to thank the final member of my committee, Richard Karp, for his excellent comments and suggestions.

A few people here have contributed directly to this work, including Kathryn Muratore, whose thesis was on the aminotransferases (and she let me pipette), Jack Kirsch, who asked incredibly insightful questions leading to new directions of research, John Srouji, who did the manual literature search of the Nudix proteins and reconstructed the GO DAG for the hydrolase family, Jason Stajich, who helped me immensely with the fungal genome work (including providing all of the genomes with the hypothetical

proteins), and Philip Johnson, who wrote a lot of the initial scripts for SIFTER and helped with its development.

I have been fortunate enough to work with or have contact with a number of insightful professors at Berkeley, in particular Kimmen Sjolander, but also Terry Speed, Ian Holmes, and Mike Eisen. I have enjoyed interactions with a large number of colleagues who have each played a role in my education, including Francis Bach, David Blei, Anat Caspi, Pat Flaherty, Ed Green, Emma Hill, Liana Lareau, Brian Milch, Andrew Ng, Xuanlong Nguyen, Guillaume Obozinski, Mark Paskin, Martin Wainwright, and Alice Zheng.

I would like to thank my sources of funding, including the National Science Foundation Graduate Fellowship and the Google Anita Borg Scholarship. I also thank Google Research, and in particular David Pablo Cohn, for hosting me for an enriching summer internship.

Finally, I thank my family and friends for their support. I especially want to thank my mother, who can explain this thesis better than I can without knowing what any of the words mean, and my father, for making me explain it to him again even though he understands it better than I do. They are steadfast, tireless allies in completely complementary ways, and without them I would be an empty vessel. Finally, I would like to thank Lance Martin for his love, patience, and kindness, and our son, Wolf Martin, the most joyful boy in the world. Their love is my sustenance.

Dedicated to my father, Dean Lee Engelhardt, Ph.D.

Contents

Chapter 1

Introduction to Phylogenomics

1.1 The promise of phylogenomics

One of the most challenging problems of central importance in the post-genome era is the prediction of protein molecular function. Challenging problems in computational biology such as this require the integration of an informed biological framework, powerful bioinformatics tools, and high-quality experimental data. In recent years, new insights into the diverse biological processes underlying the evolution of protein function have provided a powerful framework for automating protein function prediction. *Phylogenomics* [Eisen, 1998] formalizes how assumptions of molecular function evolution can be exploited to improve function prediction and to dramatically reduce the systematic errors from pairwise annotation transfer methods, the standard protocol for protein functional annotation [Sjölander, 2004; Brown and Sjolander, 2006]. This area at the intersection of phylogenetics and genomics includes methodological contributions from diverse backgrounds; it extends beyond the specific focus of this thesis, which is molecular function prediction, to include phylogenomic approaches to reconstructing species evolution, predicting features of protein ter-

tiary structure, inferring the cellular localization and biochemical pathways of extant and ancestral genes, and, moreover, reconstructing the evolutionary mechanisms by which novelty arose for each of these protein features [Thornton and LaSalle, 2000; Delsuc *et al.*, 2005]. In this introduction, we review the fundamental evolutionary processes underlying protein family functional diversification and the computational methods available for inferring these evolutionary events. We identify the challenges for functional phylogenomics in order to motivate the main aim of this dissertation, which is a statistical method for functional phylogenomics.

1.1.1 Annotation transfer

One standard approach to protein molecular function prediction employs annotation transfer, assigning the function of a characterized protein to a query protein whose function is unknown based on a significant score in database search. In this approach, homologs to the query protein are identified by searching a sequence database (typically using a computationally efficient approach such as the Basic Local Alignment Search Tool (BLAST) [Altschul *et al.*, 1990]). If the scores of the top annotated hits are significant, the implied *homology*, or evolutionary relationship, enables the annotations to be transferred directly to the query protein. This protocol is broadly used because it is straightforward to justify, implement, and run. Unfortunately, it is known to be prone to systematic error. The fundamental assumptions of annotation transfer are that sequence similarity implies an evolutionary relationship, and that evolution conserves function; biologists then infer that two proteins sharing statistically significant sequence similarity share a common function.

Many factors confound this apparently straightforward methodology. The most significant database hit may have a different function due to single point mutations, gene duplication, or domain shuffling [Eisen, 1998; Bork and Koonin, 1998; Galperin

and Koonin, 1998]. The most significant database hit may also be misannotated, and transferring this incorrect annotation propagates the annotation errors [Brenner, 1999; Devos and Valencia, 2000]. Furthermore, sequence similarity may be due to *parallel evolution* – evolution towards the same sequence from evolutionarily unrelated proteins – rather than an evolutionary relationship, especially if the significance of the hit is questionable. As many authors have pointed out (e.g., [Reeck *et al.*, 1987]), two proteins are either homologous, meaning related by a common ancestor, or they are not. There is no single E-value cutoff to describe where homology ends and non-homology begins. For a particular query protein, such a cutoff may not exist: a search may rank non-homologs higher than homologs at certain places in the search results.

Furthermore, annotation transfer based on pairwise sequence similarity does not take into account rate variation. If we assume that function generally evolves parsimoniously within a phylogeny (where we are defining *parsimonious* as each function being active within all of the proteins descendant from a single common ancestor), then the branching structure of a phylogeny is more relevant for annotation transfer than path length within the phylogeny. But lineage-specific rate variation, which is complex phenomenon prevalent across a wide range of protein families [Thomas *et al.*, 2006], means that the most similar sequences according to BLAST (i.e., those with the shortest path length in the tree) are not necessarily most likely to share a common function. One such scenario is illustrated in Figure 1.1, where path length and branching order from a query protein rank the remaining proteins differently because the phylogeny exhibits lineage-specific rate variation. This problem grows worse as more proteins are added to a tree. Specifically, additional proteins that are siblings of protein $B1$ (in the figure) with a similar branch length will provide increasingly strong support for the incorrect function transfer to the query protein. This means that the BLAST most significant hits approach is systematically flawed and may yield increasingly erroneous results as data increase. Use of a phylogeny explicitly incorpo-

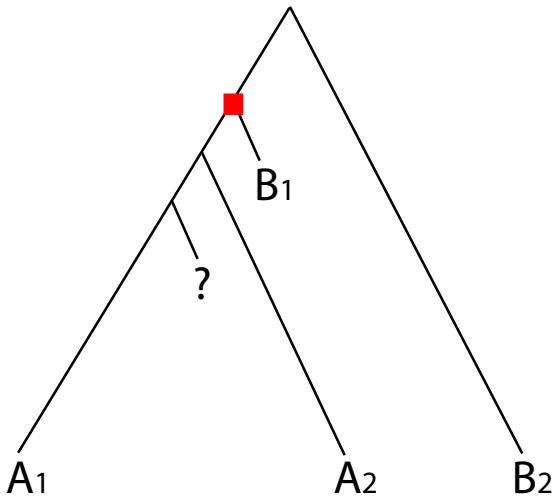

Figure 1.1: **Sequence similarity does not directly reflect phylogeny.** A phylogenetic tree shows where sequence similarity measures, such as BLAST fail to make correct functional assignments. The proteins in this tree are either molecular function A or B, with a duplication event indicated by a red square. We wish to predict the function of the protein denoted by "?". Its most significant BLAST hit will be B_1, because the path length in the phylogeny from the query protein is the shortest. Thus BLAST will transfer annotation B to the unannotated protein. However, this conflicts with the most likely molecular function scenario. It is more likely that the tree has only one functional change, in which ancestral function B mutated to function A on the left-hand side of the bifurcation. Thus A is a more likely annotation for the unannotated protein. The phylogenomics approach reaches this conclusion naturally. Adapted from [Eisen, 1998].

rates the evolutionary history and minimizes problems due to rate variation, because the branching structure is considered, instead of just branch lengths.

1.1.2 Phylogenomic analysis is designed to address these issues

Phylogenomic inference of molecular function is actually an annotation transfer protocol, but it avoids many of the errors of standard annotation transfer through the explicit inclusion of an evolutionary model of functional evolution overlaid on a phylogenetic tree. Phylogenetic methods have long been preferred to pairwise comparison methods for taxonomic data analysis (e.g., [Farris, 1982]) based on both theoretical

and practical arguments. The phylogenomic protocol is the phylogenetic approach to annotation transfer. Phylogenomics additionally relies on accurate orthology analysis, which also benefits from a phylogenetic approach.

Phylogenomics allows biologists to exploit the variable evolutionary persistence of different types of protein functional attributes to provide nuanced predictions of protein function. We use the term *evolutionary persistence* to reflect the variable endurance of protein attributes over different evolutionary distances. For instance, the structure, or three-dimensional shape, of a protein is maintained over long evolutionary distances between homologous proteins: two proteins can have vanishingly low sequence identity and still have obviously similar three-dimensional structures [Krissinel, 2007]. Basic chemical function is often maintained over long evolutionary distances as well, such as the Nudix hydrolases [Bessman *et al.*, 1996], as are general mechanisms of enzyme catalysis [Glasner *et al.*, 2007]. On the other hand, the precise biochemical function of a protein can be significantly affected by a small number of mutations. Mutations at an enzyme catalytic site can completely disrupt function, while mutations at nearby binding pocket positions may modify the substrate upon which the enzyme acts. As an example, in *Halobacterium salinarium*, bacteriorhodopsin (light-driven proton pumps) and halorhodopsin (chloride ion pumps) are homologous; mutation of an aspartic acid to a threonine changes this light-driven proton pump to a chloride ion pump [Sasaki *et al.*, 1995]. Cellular localization, like structure, has a long evolutionary persistence in the absence of domain shuffling [Chen and Rost, 2002]. Like molecular function, small evolutionary changes can drastically impact particular localizations. For example, a protein that is secreted from the cell may change to one that is cytoplasmic based on the presence or absence of short signal peptides [Tagaya *et al.*, 1997], and transmembrane localization may change to localization in the Golgi apparatus after a small number of residue changes [Wong *et al.*, 1992]. Phylogenomics takes into account the variable

evolutionary persistence of different protein features to enable a more precise and nuanced prediction of those features than a simple pairwise comparison of individual proteins.

The basic procedure for a phylogenomic analysis of a *protein superfamily*, or a family of proteins generated from a common ancestor by gene duplication and speciation events, is as follows (as in Figure 1.2): Starting with a query sequence, homologous sequences are gathered according to some criteria. A multiple sequence alignment is constructed for these sequences and is used as the basis for estimating a protein superfamily phylogeny. Reconciliation of the rooted superfamily tree and a trusted phylogeny of the species is then used to localize gene duplication and speciation events at the internal nodes of the superfamily phylogeny [Goodman *et al.*, 1979; Page, 1998]. These events enable identification of subtrees in the protein superfamily phylogeny that correspond to orthologous proteins (proteins related by speciation events). Molecular function annotations derived from an experiment are then overlaid on the tree, highlighting functional shifts within the superfamily. The reconciled tree including molecular function annotations can then be used as the basis of annotation transfer. This introductory chapter will serve to motivate this general protocol. The remainder of the dissertation will introduce a particular method for inferring molecular function from a reconciled tree overlaid with molecular function annotations, and show results and additional methods that can be used in conjunction with this phylogenomic protocol.

1.2 Overview of protein superfamily evolution

The basic events in protein family evolution include single nucleic acid changes (including mutation, insertion, and deletion), gene duplication and deletion, and domain shuffling (including repetition, insertion, deletion, and exchange of protein domains).

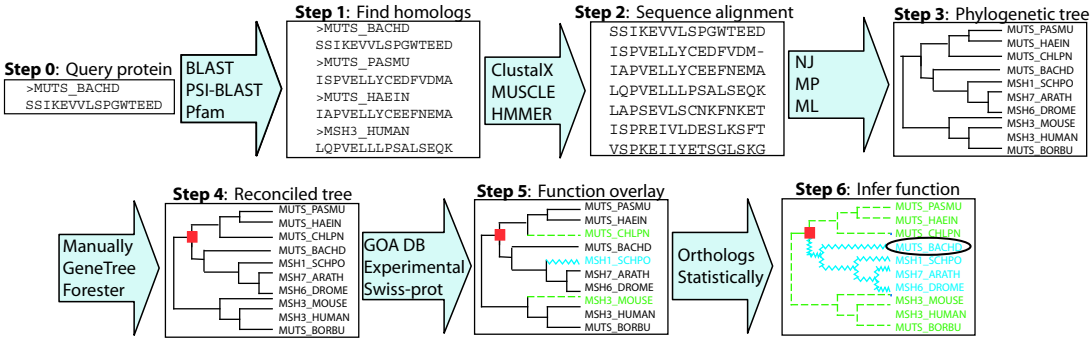

Figure 1.2: **The basic pipeline for phylogenomic inference of molecular function.** Beginning with a query protein sequence, the phylogenomic approach builds a reconciled tree and overlays molecular function annotations to infer the function of the query protein.

Speciation and gene duplication events are the two branching processes that generate protein superfamilies in combination. A few terms developed elsewhere are helpful in categorizing relationships within protein superfamilies, and we define and illustrate each of them in Figure 1.3. Proteins that are related by duplication events at their most recent common ancestor are *paralogs*, whereas proteins related by speciation events at their most recent common ancestor are *orthologs*. Zmasek and Eddy [Zmasek and Eddy, 2002] define additional terms based on analysis of a rooted, strictly bifurcating tree. *Ultra-paralogs* are proteins in the same species that have only duplication events on the internal nodes of their direct tree path. *Super-orthologs* are two proteins in different species that have only speciation events on the internal nodes of their direct tree path. *Subtree-neighbors* of order k are all descendants other than the protein itself of the closest k ancestors of a protein (where k is generally 2). In general application, subtree-neighbors can be used as the basis of annotation transfer across members of a clade, assuming consistent annotations. O'Brien and colleagues [O'Brien *et al.*, 2005] further define *inparalogs* as proteins in the same species that are most recently related through a duplication event within that species, and

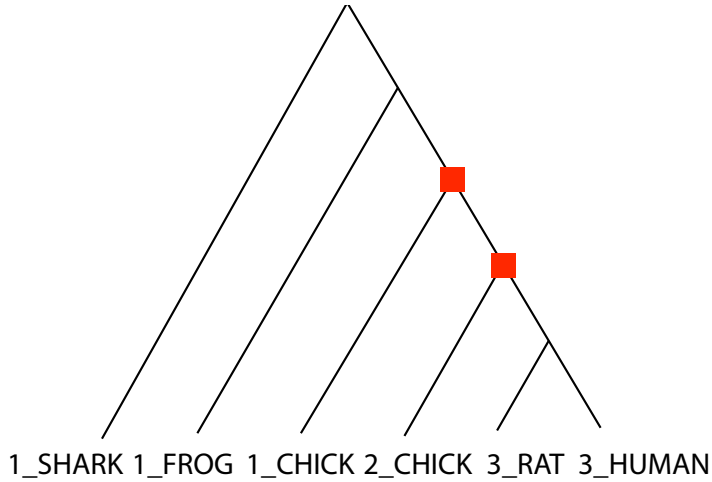

Figure 1.3: **Homology relationships defined.** In this phylogeny, the red squares represent duplication events, and the proteins are named by their copy number followed by a species. Proteins 1_CHICK and 2_CHICK are paralogs, as are proteins 2_CHICK and 3_HUMAN. Proteins 3_RAT and 3_HUMAN are orthologs, and so are proteins 3_RAT and 1_FROG. Proteins 2_CHICK and 1_CHICK are the only ultra-paralogs in the figure. Proteins 3_RAT and 3_HUMAN are super-orthologs, as are proteins 1_SHARK and 1_FROG. Proteins 3_RAT and 3_HUMAN are subtree-neighbors of protein 2_CHICK, as is protein 1_CHICK. Proteins 1_CHICK and 2_CHICK are inparalogs, and proteins 3_RAT and 2_CHICK are outparalogs (as are proteins 3_HUMAN and 2_CHICK).

outparalogs as proteins in different species related via a duplication event at their most recent common ancestor, that have more recently undergone a speciation event (and perhaps additional duplication events). These definitions will help in the discussion of orthology.

1.3 Gene duplication and orthology

The inference of orthology is of great interest in functional genomics due to the assumption that orthologs are likely to share a common function. This has resulted in the development of databases of predicted orthologs or orthologous groups of proteins (e.g., COGs [Tatusov *et al.*, 2000], TOGA [Lee *et al.*, 2002], InParanoid [O'Brien

et al., 2005]). Gene duplication provides a central evolutionary mechanism enabling functional innovations in protein superfamilies. Following gene duplication, selective pressures on the original copy of a gene can diminish; one or both of the proteins may modify their molecular function, either as a result of neutral evolutionary drift or selection. The mechanistic details of gene duplication, however, are quite complex and poorly understood, and much recent research has attempted to extrapolate these details from the examples of gene duplication events.

1.3.1 Evolution of molecular function

Sequence evolution, on one hand, is constrained by a protein's role in the survival and fitness of an organism, and on the other hand is subject to forces of evolutionary change, including mutation, duplication, and deletion events. In order to understand better this evolutionary dynamic, we briefly review the current literature on molecular evolution as it relates to molecular function and evolutionary constraints. The terms *gene* and *protein* (i.e., gene product) are used somewhat interchangeably, ignoring some subtleties that do not broadly impact this discussion.

Although the rate of molecular evolution does not directly determine the rate of molecular function evolution, the phylogenomic assumption that sequences and molecular functions evolve at proportional rates will guide our discussion. In particular, although much work has been done on the rates of molecular evolution for particular proteins (and many open questions still exist), much less work has been done on the rates of molecular function evolution. Generally, we wish to identify which characteristics of proteins (and the evolutionary history of these proteins) lead to a faster rate of molecular function evolution. This will enable a model for the approximate localization of functional mutations in the phylogenetic history of a set of homologous proteins.

1.3.1.1 Essentiality, dispensibility, and expression levels may correlate with a slower rate of evolution

Researchers have attempted to find correlations between the evolutionary rate of proteins (as a measure of how strong evolutionary constraints are on this protein) and different features of the protein *in vivo*. One such feature is protein *essentiality*, which describes whether or not a protein is critical to the survival and viability of the organism. The *knockout rate hypothesis* negatively correlates the impact of knockout on the fitness of the organism with rate of evolution [Wilson *et al.*, 1977]. Eukaryotic genes that are essential to the viability of an organism appeared to be more constrained by selection, and also appeared to arise farther back in evolutionary time [Decottignies *et al.*, 2003], as observed in a large-scale knockout study of genes in *S. pombe* and *S. cerevisiae*. This general hypothesis was challenged in a study in mouse and rats, which showed that sequence mutation rates were not correlated with essentiality of the gene [Hurst and Smith, 1999].

More sophisticated studies have measured the impact of gene *dispensibility* on mutation rates for that gene, where dispensibility is often measured by growth rates of an organism with the particular gene deleted. The general idea is that, for genes that are not essential to the organism's survival but contribute to organismal fitness, most mutations would be considered within the range of neutral. Thus, because they have fewer selective constraints than essential genes, the rate of evolution should be higher in genes that contribute less to the overall fitness of the organism. This hypothesis is hard to test directly because the fitness of a particular protein to an organism is difficult to quantify.

Hirsh and Fraser found a significant correlation between protein dispensibility and evolutionary rate, and note in their results that gene essentiality has not been definitively correlated with evolutionary rate (in their own measurements in yeast

and in the previously cited mouse study [Hurst and Smith, 1999]) because the distinction between essential and non-essential genes was not precise enough [Hirsh and Fraser, 2001]. In particular, when a gene knock-out results in a very slow rate of organismal growth, the individual has such poor fitness that evolutionarily the gene can be considered almost equivalent to essential. Yang and colleagues followed this up with a similar experiment in yeast and *C. albicans* that found a weak correlation between dispensibility and evolutionary rate, but noted that the correlation disappeared when duplicate genes were not considered [Yang *et al.*, 2003a]. They also hypothesize that structural constraints influence the rate of evolution more than dispensibility. Wall and colleagues further corroborated these findings in a study that found evidence for independent correlations between the rate of molecular evolution and both gene dispensibility and expression levels [Wall *et al.*, 2005]. Although they argue based on their own findings that dispensibility is correlated with the rate of evolution, they cite two previous studies ([Pal *et al.*, 2003; Rocha and Danchin, 2004]) that both conclude that the correlation between dispensibility and the rate of evolution is not significant when expression level (or, as in the second study, expression levels and molecular function type) is taken into account.

1.3.1.2 Pleiotropy constrains evolutionary rate

In 1930, Fisher proposed that pleiotropy constrains evolution [Fisher, 1930]. *Pleiotropy* means that a single gene is responsible for multiple phenotypes. It could be that a single gene has multiple molecular functions, is involved in multiple pathways, or interacts with different proteins depending on tissue type.

Much more recent studies of the yeast protein interaction network have shown correlations between the number of proteins interactions for a given protein and evolutionary constraints. One such study found a correlation between number of protein interactions with decreased viability of the organism after knockout [Jeong *et al.*,

2001]. Another study showed that a protein that is more central in a genetic pathway tends to be more critical and have a slower rate of evolution [Hahn *et al.*, 2005]. Krylov introduced the propensity for gene loss (PGL) measure, which is computed by comparing the phylogenetic profile to the species phylogeny, and showed that PGL is more negatively correlated with gene essentiality (as measured by knockout analyses) and number of interaction partners than the rate of sequence evolution [Krylov *et al.*, 2003].

Returning again to Fisher's hypothesis, proteins involved in more protein-protein interactions have been shown to have a slower evolutionary rate not because of their essentiality to the organism, but because a larger proportion of amino acids in the protein are involved in functional interactions, thus those amino acids will evolve at a slower rate [Fraser *et al.*, 2002]. A specific amino acid might be involved in a functional interaction either directly, through involvement in direct bonds with functional molecules, or indirectly, through their role in protein structural stability. As the number of interactions of a particular protein grows, so does the number of sites that are explicitly constrained.

1.3.1.3 Gene duplicability

The evolutionary role of gene duplication, beyond creating novel proteins, still remains unclear. One alternative role might be as a compensating mechanism for gene deletion. In yeast single-gene-deletion mutants, one study estimated that one quarter of the deletions showed no phenotype because of compensation from a duplicated gene in the genome [Gu, 2003]. Furthermore, as the sequence similarity between gene duplicates decreases, their ability to compensate for a deleted copy decreases, presumably because the function diverges. The surprising result, though, is that even evolutionarily distant gene duplicates (i.e., those that do not obviously have the same molecular function) may compensate for a deleted gene.

One alternative way to explain these data is that critical genes are duplicated at a much lower frequency. Gene *duplicability* is defined as the tendency of a gene to duplicate and subsequently be fixed in that species. This hypothesis implies that what appears to be gene compensation is actually just a correlation between non-critical genes and duplications, and was supported in a set of yeast gene knockout studies [He and Zhang, 2006]. This hypothesis came from two observations. First, genes involved in large protein complexes have lower duplicability because changing (e.g., doubling) the concentration of the molecules involved in the interactions may detrimentally impact the functional products of the interaction network [Jeong *et al.*, 2001]. There is substantial evidence to suggest that proteins involved in more interactions will be, on average, more critical to organism survival as measured by knockout phenotype [Yang *et al.*, 2003b].

A final study in this vein postulates that when genes that are more biologically important duplicate, they are retained at a much higher rate than duplicates of genes that are less biologically important in eukaryotes [Jordan, 2004]. This observation comes from evidence that, despite an initial increase in evolutionary rate following a duplication, the evolutionary rate of the retained duplicate genes dramatically slows, possibly because of the high number of selective constraints on these more important proteins.

This leaves us with a number of open questions about how gene duplication relates to pleiotropy, biological importance, and evolutionary rate. Many details of these correlations still remain to be resolved among the studies presented here in order to understand better the details of how duplication contributes to functional diversity in proteins. In many cases, it is not even clear how hypotheses can be tested given our limited data and indirect observation of duplication.

1.3.2 Gene duplication events

There are several scenarios of how gene duplication events lead to functional divergence, illustrated in Figure 1.4. The majority of gene duplications result in *nonfunctionalization*, where one copy of the gene acquires neutral mutations, converting it into a pseudogene. *Neofunctionalization*, proposed by Ohno [Ohno, 1972], hypothesizes that, after a duplication event, one copy of the gene maintains the original function, while the other acquires a novel, adaptive function through positive selection. One example of neofunctionalization is the sensory-neuron-specific (SNSR) subclass of G-protein coupled receptors (GPCRs), in which researchers have observed that the ligand-binding residues underwent positive selection following gene duplication resulting in functional divergence of the paralogs [Choi and Lahn, 2003].

An alternative theory, known as *subfunctionalization*, or the duplication-degeneration-complementation (DDC) model, postulates that mutations accumulate via neutral evolution in each copy of a gene, resulting in complementary functions [Force *et al.*, 1999; Lynch and Force, 2000]. Two subfunctionalization scenarios exist. In the first scenario, specifically the DDC model, complementary functions are lost in the duplicate genes. One example is the sex-linked $\alpha 4$ proteasome subunit genes in *Drosophila melanogaster*, in which two copies of the gene, both expressed at different times during spermatogenesis, are each missing different functional regions [Torgerson and Singh, 2004]. In the second subfunctionalization scenario, a protein with a general enzymatic function, possibly working on a large range of substrates, develops complementary substrate specificity in the duplicate genes. In corn, phytochrome genes were observed to have developed overlapping but differentiated functions due to this type of subfunctionalization [Sheehan *et al.*, 2007]. For more information on this topic, many good reviews of gene duplication exist (e.g., [Roth *et al.*, 2007; Conrad and Antonarakis, 2007]). These scenarios all lead to the conclusion that

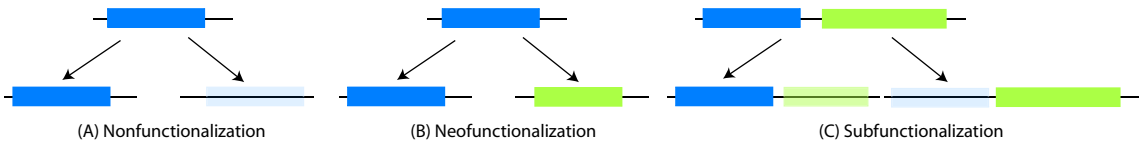

Figure 1.4: **Three different theories of functional diversification following a gene duplication event.** *Nonfunctionalization*, shown in (A), is the most biologically common of the three events, in which a duplicated gene's function is lost by neutral mutations. *Neofunctionalization*, shown in (B), shows one copy evolving an independent function. *Subfunctionalization*, shown in (C), shows the duplicated genes losing complementary functions.

changes in molecular function will often co-occur with duplication events when both copies of the gene are retained as expressed proteins.

1.3.3 Evolutionary events modify domain architecture

Domain shuffling, illustrated in Figure 1.5, is another evolutionary mechanism for functional mutation that played a significant role in the evolution of eukaryotic organisms [Liu and Grigoriev, 2004; Babushok *et al.*, 2007]. Phylogenetic studies in protein families with domain shuffling, such as the DNA-binding domain KilA-N found in bacterial and eukaryotic DNA viruses [Iyer *et al.*, 2002], show the extent of rearrangement of homologous domains in these families and the impact on domain architecture and molecular function. In addition, there are two complementary evolutionary mechanisms, *gene fusion events*, which fuse two separate genes or domains, and *gene fission events*, which split a single gene into two genes, illustrated in Figure 1.6.

The NitFhit protein in *D. melanogaster* and *C. elegans*, for example, is a histidine triad homolog fused with a nitrilase homolog; evidence suggests that the two domains may have fused because they both have nonessential functions and are involved in the same signaling pathway [Semba *et al.*, 2006]. Domain shuffling and gene fission/fusion events generate proteins with a different domain structure than the original protein,

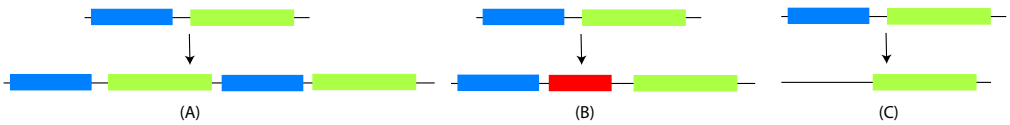

Figure 1.5: **Domain shuffling events.** Four events result in changing protein domain structure: exchange, repetition, insertion, and deletion (exchange is not shown here). Panel (A) shows repetition leading to a different domain structure in the resulting protein. Panel (B) shows a new (red) domain inserted between the two domains of the original protein. Panel (C) displays the result of deletion of the blue domain. Although this figure shows shuffling events respecting domain boundaries, this is often not the case; furthermore, the resulting hybrid genes may be modified by subsequent gene fusion or fission events. These four events in combination have the effect of producing the full range of protein domain structures.

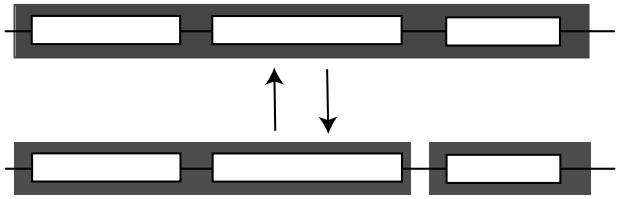

Figure 1.6: **Gene fission/fusion events.** The grey background represents the expressed protein boundaries. The up arrow represents a gene fusion event, where two separate proteins are fused into a single gene product; the down arrow represents a gene fission event, where a single protein with three domains is split into two separate proteins.

modifying the tertiary structure of the protein and often the molecular function, sometimes fairly dramatically [Pasek *et al.*, 2006]. Although it is clear that domain shuffling and gene fusion/fission events play a role in the evolution of molecular function [Ashby and Houmard, 2006; Alm *et al.*, 2006] and protein networks [Amoutzias *et al.*, 2004], most automated phylogenomic methods do not explicitly account for these evolutionary mechanisms. Both domain shuffling and gene fission and fusion result from recombination events [Kummerfeld *et al.*, 2004].

1.4 Fundamental assumptions of phylogenomic analysis

Two assumptions about protein molecular function evolution are of critical importance for phylogenomic analysis. The first critical assumption is that molecular function evolves in parallel with protein sequence. In other words, the phylogeny reconstructed for a set of homologous molecules based on sequence information is a satisfactory proxy for the evolution of molecular function and structure in those proteins. The second critical assumption is that discrete changes in molecular function most often co-occur with gene duplication events. These two assumptions allow us to develop a model of molecular function evolution to generate experimentally testable hypotheses.

The two assumptions should be examined carefully. The first assumption requires (a) the availability of an accurate phylogenetic tree reconstruction for a protein superfamily, and (b) that the inferred branches of this phylogeny along which the greatest change in protein sequence occurred are, in general, the same branches along which the greatest change in protein molecular function occurred (taking into account the locations of duplication events). We will defer examination of (a) until the discussion of tree reconstruction methods; (b), however, is only a general tendency, which is often violated for particular superfamilies. For example, single nucleotide mutations may produce discrete changes in molecular function, explicitly violating this assumption, as in the case of the lactate dehydrogenase protein from *T. vaginalis*, evolved from a malate dehydrogenase, where a single amino acid mutation is thought to be responsible for the functional change [Madern, 2002; Wu *et al.*, 1999]. *Parallel functional evolution*, or a single molecular function which arises independently in different locations in the phylogeny (also a less common definition of *convergent evolution*), although it violates an assumption of maximally

parsimonious evolution, does not explicitly violate this assumption.

Although we are unaware of any genome-scale studies comparing the rate of functional change after gene duplication versus after speciation events, one study found that, for a particular functional type (i.e., defense genes), the function itself was more determinant of a higher rate of evolution than the existence of a paralogous gene in humans [Nembaware *et al.*, 2002]. This might indicate that the factors contributing to duplication of a protein (i.e., its duplicability), such as the molecular function and the length of the protein, may more directly impact functional change than duplication itself. These hypotheses imply a modified phylogenomic methodology that uses additional protein features that impact duplicability, including length of a protein sequence and the type of functions being inferred, to more accurately estimate evolutionary persistence of function within a multigene phylogeny.

1.5 A simple phylogenomic analysis

We outline the basic steps involved in a phylogenomic analysis of a set of protein sequences in Figure 1.2, given the basic assumptions outlined above. We assume that the process is initiated with a query protein sequence of unknown function. We motivate each step, and describe methods that may be used and how those methods perform in practice. We focus on a few potential sources of noise or error that may be relevant to a phylogenomic analysis, areas for improvement, and open research questions.

1.5.1 Step 1: Identify a set of homologous proteins

1.5.1.1 Motivation, definition, and methods

We would like to gather a set of proteins homologous to the query protein in order to transfer a functional annotation according to the phylogenomic protocol. The selection of sequences in the first step of a phylogenomic analysis has a significant impact on the resulting functional inferences. There are three basic approaches to this task: selecting sequences based on individual alignments of database hits to the sequences of interest, selecting homologous sequences from databases that cluster homologous proteins (including Pfam [Bateman *et al.*, 2002], SCOP [Andreeva *et al.*, 2004], COG [Tatusov *et al.*, 2000]), and selecting sequences found to have the same domain architecture as the sequences of interest. Standard homolog clustering methods (e.g., BLAST [Altschul *et al.*, 1990] and PSI-BLAST [Altschul *et al.*, 1997]) are optimized for the detection of local matches to a query sequence. PSI-BLAST constructs a profile HMM from a multiple sequence alignment of sequences with significant BLAST scores to detect remote homologs, but including non-homologs (or partial homologs) accidentally can generalize the profile so that non-homologous sequences are found using the profile. The homology databases often rely on these sequence search methods (e.g,. COGs uses triangles of mutually-consistent, genome-specific best hits from BLAST) for populating their clusters.

One might also restrict sequences to those that share a common domain architecture with the query sequence, when these sequences include a broad enough sampling of experimentally-annotated sequences with homology at functional regions of interest. The programs FlowerPower and CDART are designed specifically to cluster sequences having the same domain architecture [Krishnamurthy *et al.*, 2007; Geer *et al.*, 2002]. FlowerPower is similar to PSI-BLAST [Altschul *et al.*, 1997], but instead of using a single profile to expand the existing cluster, FlowerPower uses a set

of HMMs: a general HMM for the family as a whole, and a subfamily HMM (SHMM) for each predicted subfamily. The Conserved Domain Architecture Retrieval Tool (CDART) also performs searches for proteins with identical domain architecture [Geer *et al.*, 2002] using Reverse PSI-BLAST (RPS-BLAST) to search domain profiles built from the Conserved Domain Database [Marchler-Bauer *et al.*, 2002], with additional processing to handle redundant or closely-related domains.

1.5.1.2 Limitations and considerations

In selecting homologs, investigators must strike a balance between a conservative selection of close homologs and retrieval of more divergently related sequences. Restricting the set to close homologs with common domain architecture can produce good alignments and accurate tree topologies, but may result in insufficient information for function inference due to the sparse nature of experimental data. Including distant homologs and more diverse domain architecture can increase the available experimental data, but can lead to errors in alignment and tree topology, and result in incorrect or overly general functional predictions. Functional inferences based on distant homologs must be made with caution, as estimation of evolutionary persistence is difficult over long evolutionary distances.

1.5.2 Step 2: Align the sequences

1.5.2.1 Motivation, definition, and methods

The multiple sequence alignment (MSA) is the source of phylogenetic signal, and so plays an important role in phylogenomic analysis. Numerous multiple sequence alignment packages are available that are both computationally efficient and produce high-quality alignments. These include MUSCLE [Edgar, 2004] and MAFFT [Katoh *et al.*, 2002] for large multigene families, and PROBCONS [Do *et al.*, 2005] for smaller

families. All have produced outstanding results on benchmark datasets comparing sequence and structural alignments [Edgar and Batzoglou, 2006]. Alternatively, given a hidden Markov model (HMM) profile for this particular protein or domain, HMM alignment methods such as `hmmalign` [Eddy, 1998] can be used to quickly align the sequences given the profile. Post-processing of the MSA may further improve the signal, and includes alignment masking, deleting poorly-alignable sequences, using three-dimensional structure to improve the alignment, or cropping an alignment to a conserved core.

1.5.2.2 Limitations and considerations

There are dramatic differences between the characteristics of single-gene (orthologous) groups and groups of genes related by gene duplication (multi-gene families), particularly when these groups span large taxonomic distances. The former often have high sequence identity, and their phylogeny will resemble the species phylogeny. The latter often contain divergent sequences with low sequence identity (e.g., some pairs may not be more identifiably similar than two random sequences), regions where the alignments are unreliable and the proteins may not even be structurally superposable, or extreme variability in site- and lineage-specific mutation rates [Saier, 1996]. All of these protein superfamily features strain the core assumptions of multiple sequence alignment methods, which attempt to create a sequence of columns that are each evolved from a common ancestor amino acid. This implies that, superposing the three-dimensional structures from this family, each position of the superposed complex is a single column of the alignment. When the structures cannot be superposed, it is possible that these assumptions are violated.

To address this issue, masking is often employed. *Alignment masking* removes columns in the multiple sequence alignment that appear to be uninformative. Often in masking, columns of a multiple sequence alignment are removed when, for

example, greater than 75% of the column is gaps instead of informative characters. However, this practice removes information that may be important in reconstructing phylogenetic lineages. For instance, sequence motifs that uniquely identify family subtypes may be targeted for masking prior to phylogenetic tree reconstruction, but their functional importance and evolutionary signal to differentiate subtypes could be exploited in phylogenetic tree reconstruction.

1.5.3 Step 3: Reconstruct a phylogeny

1.5.3.1 Motivation, definition, and methods

The choice of phylogenetic reconstruction method depends on the computational resources available and the size of the dataset to be analyzed. Distance methods (e.g., UPGMA and neighbor-joining (NJ) [Felsenstein, 1989]) are widely used when computational efficiency is an issue, as in construction of phylogenies for each protein in a genome. Many different variants of the standard NJ protocol have been developed into programs and evaluated on a range of protein families [Hollich *et al.*, 2005]. Character methods include maximum parsimony (MP), maximum likelihood (ML), and Bayesian methods. MP methods are significantly slower to reconstruct than the distance methods, and are often less accurate than standard distance methods using a high-quality distance metric (e.g., [Atteson, 1997; Felsenstein, 1978]). It has been noted that MP methods perform particularly poorly when there are long branches in the phylogeny [Felsenstein, 1978]. ML methods are more accurate than both distance methods and MP methods on simulation studies [Kuhner and Felsenstein, 1994; Tateno *et al.*, 1994], but are also computationally slower than both. The computationally slowest algorithms for phylogeny reconstruction are the Bayesian methods, and although they may confer a small performance advantage over ML methods, their applicability is limited to very small families [Mar *et al.*, 2005; Hall, 2005].

The PHYLIP [Felsenstein, 1989] resource includes a compilation of numerous tools for phylogenetic reconstruction. Other popular resources include PHYML [Guindon and Gascuel, 2003], which is a recent maximum likelihood tree reconstruction method designed for computational efficiency, PAUP* [Swofford, 2001], which includes MP, ML, distance methods, and bootstrapping capabilities, and MrBayes [Huelsenbeck and Ronquist, 2001], which takes a Bayesian approach to tree construction, performing a Markov chain Monte Carlo search through the space of possible phylogenies.

Rooting a phylogeny, or identifying the branch that contains the most recent common ancestor of each of the phylogeny leaves (which implicitly makes each branch directed) is required to localize gene duplication events, but many phylogenetic tree construction programs produce unrooted trees. Selecting outgroup sequences in a phylogenetic reconstruction is standard practice for rooting species phylogenies, but is often not possible in *multi-gene families* (i.e., genes from outgroup species may appear in multiple subtrees in the reconstructed phylogeny, or the root may be an ancient duplication event making it difficult to identify an outgroup) [Felsenstein, 2003]. *Mid-point rooting* places the root at the midpoint of the longest span in the tree and can be applied to multi-gene families [Felsenstein, 2003]. This approach assumes a molecular clock, which may not be a reasonable assumption in many protein superfamilies due to the potential for lineage-specific rate variation following gene duplication. A related method with the same assumption places the root to minimize the difference in path lengths from the root to terminal nodes across the tree [Felsenstein, 2003]. *Parsimony-based rooting* places the root of an unrooted gene tree so as to minimize the number of gene duplications and gene losses in the tree [Berglund-Sonnhammer *et al.*, 2006; Thornton and LaSalle, 2000]. This is method appears to be more robust to lineage-specific rate variation, but assumes that data being analyzed come from fully sequenced genomes. In contrast, mid-point rooting can be applied to phylogenomic analysis of sequences retrieved from partially sequenced genomes.

Another type of method that is often employed computes statistical measures of subtree reliability. The measurement should reflect how evidence in the MSA supports this particular evolutionary topology. Bootstrap analysis can be used to estimate the support for subtrees [Felsenstein, 1985]. Measures based on the bootstrap, including the approximately unbiased test [Shimodaira, 2002], are often more powerful (and more easily implemented) than the standard bootstrap in evaluating overall phylogeny confidence. There are many possible alternative probabilistic measures of subtree reliability, including posterior probabilities based on Bayesian analysis [Huelsenbeck, 1995] or approximate likelihood ratio tests [Anisimova and Gascuel, 2006]. These measures can be used to estimate overall tree reliability, and may be helpful to factor into the final phylogenomics-based predictions, perhaps to down-weight predictions based on an inaccurate subtree as in Orthostrapper [Storm and Sonnhammer, 2002] and RIO [Zmasek and Eddy, 2002].

1.5.3.2 Limitations and considerations

The quality of phylogenetic tree reconstructions is commensurate with the quality of the input multiple sequence alignment, where the problems caused by multigene families have been studied in detail. In general, phylogenetic tree reconstruction methods have been developed for reconstructing the evolutionary history of single-gene families; while their accuracy in reconstructing single-gene family phylogenies is approximately known (primarily through simulation studies, e.g., [Huelsenbeck, 1995]), their accuracy in reconstructing phylogenies of multi-gene families is not well understood.

One assumption that is strained in multi-gene families is *positional homology*, i.e., that the characters in a multiple sequence alignment column descend from a single ancestral character. Positional homology is a fundamental assumption in character based phylogenetic tree reconstruction methods, which assumes that the characters in

each column of an alignment evolved independently and identically distributed (IID). However, structural studies have shown that, as two proteins diverge from a common ancestor, their structures also diverge at a slower rate [Baker and Sali, 2001]. As evolutionary distances increase, the sequence similarity can become extremely low, even when the core structural elements are maintained. Constructing a reliable multiple sequence alignment for such divergent sequences is difficult. The lack of structural superposability across all positions in homologous proteins often violates the positional homology assumption for the multiple sequence alignments (e.g., insertions in loop regions). Detailed comparisons of alignment methods applied to divergently related sequences have shown that no alignment methods produce high-accuracy results when sequence identities fall below 30% [Baker and Sali, 2001]. These issues are not problematic in the analysis of single-gene families, where sequence identity rarely falls below 30% [Thompson *et al.*, 1999].

Phylogenetic tree reconstruction is also sensitive to the specific region of alignment. When highly divergent sequences from multi-domain protein superfamilies are aligned, phylogenetic signal can be non-uniform across the MSA. Estimating a phylogeny from each region separately can produce different topologies, supporting distinct predictions of function for members of the family. One example of this is for large-subunit ribosomal DNA, where removing any portion of the alignment results in a modified (and incorrect) phylogenetic tree topology regardless of the tree reconstruction method [Mugridge *et al.*, 2000].

A consequence of the relative sparseness of genome sequencing in phylogenetic reconstruction is that long branches (indicating a large evolutionary distance) may be incorrectly placed. Long-branch attraction is a related problem: if there are two or more rapidly evolving sequences, some phylogenetic reconstruction algorithms will characterize them as siblings rather than as long branches from more distant ancestors [Bergsten, 2005]. A more thorough sampling of taxa can sometimes alleviate

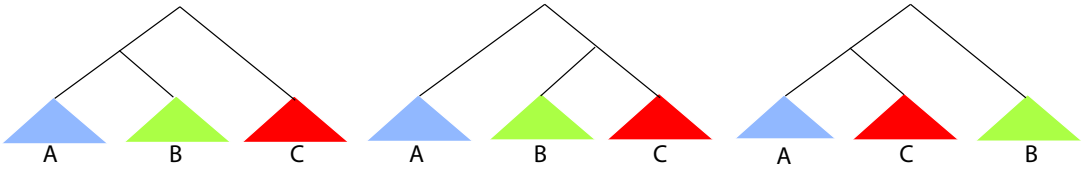

Figure 1.7: **Ambiguity in coarse branching order can influence phylogenomic inference.** Subtrees A, B and C represent orthologous groups, and are found consistently across all trees estimated for these taxa. In this illustration, subtree A does not include any sequence with experimentally determined function, but subtrees B and C have experimental data supporting distinct functions. If we use the subtree-neighbor approach to infer function, in the first tree (from left), subtree A is predicted to have a function similar to that of subtree B. In the middle subtree, it is not clear which function subtree A has. In the subtree at right, subtree A is predicted to have a function similar to C. If only one tree is included in the analysis, the ambiguity in tree topologies (and hence inferred function) will not be taken into account.

these problems, but may not be possible.

A final issue for phylogenetic tree reconstruction methods, with implications to phylogenomic inference, is ambiguity in coarse branching order. A single method may produce trees with different coarse branching orders in different runs; this is mostly because there is generally less evolutionary signal to estimate the branching order near the root of the tree. These differences in tree topologies can result in different predictions of function for subclades with no experimental functional evidence, as illustrated in Figure 1.7. For a good review of these issues and detailed recommendations see [Thornton and LaSalle, 2000].

1.5.4 Step 4: Identify duplication events

1.5.4.1 Motivation, definition, and methods

The most rigorous automated methods for localizing gene duplication events on a tree reconcile the gene tree and a reference species tree [Goodman *et al.*, 1979], such as Forester [Zmasek and Eddy, 2001a] and GeneTree [Page, 1998], as illustrated in

Figure 1.8. Approximate reconciliation, implemented in the program LOFT [van der Heijden *et al.*, 2007], examines the overlap of the species in the subtrees. In particular, an internal node will be labeled as a speciation event when the sets of species at the leaves of its (two) branches are disjoint. Although this method is quite heuristic, and is equivalent to reconciliation with a completely unresolved species tree, it does not rely on species tree accuracy.

While full phylogenetic tree reconstruction and species tree-gene tree reconciliation is the most accurate approach to orthology identification when a trusted species tree is available, it is also computationally intensive. For this reason, some methods use pairwise sequence comparisons to simply predict orthology relationships. InParanoid [Remm *et al.*, 2001] and OrthoMCL [Li *et al.*, 2003] use reciprocal best hits in BLAST to predict orthologs. InParanoid clusters orthologs (excluding inparalogs) using a rule-based approach. OrthoMCL identifies inparalogs by best reciprocal hits within a species. OrthoMCL subsequently clusters the sequences into orthologous groups using a random walk on a Markov transition matrix based on all-versus-all BLAST scores.

Other methods to construct orthologous clusters work by symbolically marginalizing out all phylogenies so that errors in tree reconstruction and reconciliation have less of an impact on ortholog determination. The result is, for each protein, a probability that it is orthologous to the query protein. For a large number of protein sequences, enumerating all of the phylogenies is not computationally feasible. Two types of approximations to this enumeration are the bootstrap approach, which sums over a set of probable phylogenies and associated reconciliations, and Markov chain Monte Carlo (MCMC) techniques, which use Bayesian sampling to marginalize over the most likely phylogenies. Softparsmap uses a reconciliation method to explicitly minimize the number of gene duplication and loss events implied by the maximum (soft) parsimony tree tree [Berglund-Sonnhammer *et al.*, 2006]. Two methods that

take the bootstrap approach, Resampled Inference of Orthologs (RIO) [Zmasek and Eddy, 2002] and Orthostrapper [Storm and Sonnhammer, 2002], are described in detail below.

There are many programs that identify putative orthologs automatically, including InParanoid [Remm *et al.*, 2001], OrthoMCL [Li *et al.*, 2003], Softparsmap [Berglund-Sonnhammer *et al.*, 2006], Forester [Zmasek and Eddy, 2001a], GeneTree [Page, 1998], and LOFT [van der Heijden *et al.*, 2007]. Although these methods are automatic, they have a high false positive rate due to many confounding factors, including the approximations employed, sensitivity to errors in the protein phylogeny (except for InParanoid and OrthoMCL, because they do not use phylogeny), and the selected set of homologs.

The methods that perform approximate reconciliation are fast relative to phylogenetic tree reconstruction methods. Due to the high rate of false positives, often manually incorporating knowledge of large-scale duplication events in the history of a collection of species (e.g., [Eisen and Hanawalt, 1999]) and identifying obvious redundant genes in sequence databases improves the accuracy of the duplication events.

1.5.4.2 Limitations and considerations

Simple sequence similarity-based orthology inference is not expected to be as reliable as inference based on full phylogenetic tree reconstruction [Searls, 2003]:

> *Tree reconciliation is the most reliable method for identifying orthologous subgroups, despite dependencies on the inherently noisy processes of multiple sequence alignment and tree reconstruction. Joint analysis of a phylogenetic tree, experimental data, alignment and structure also provides a framework for identifying the actual sequence and structural changes responsible for protein function mutations. Furthermore, orthologous clus-*

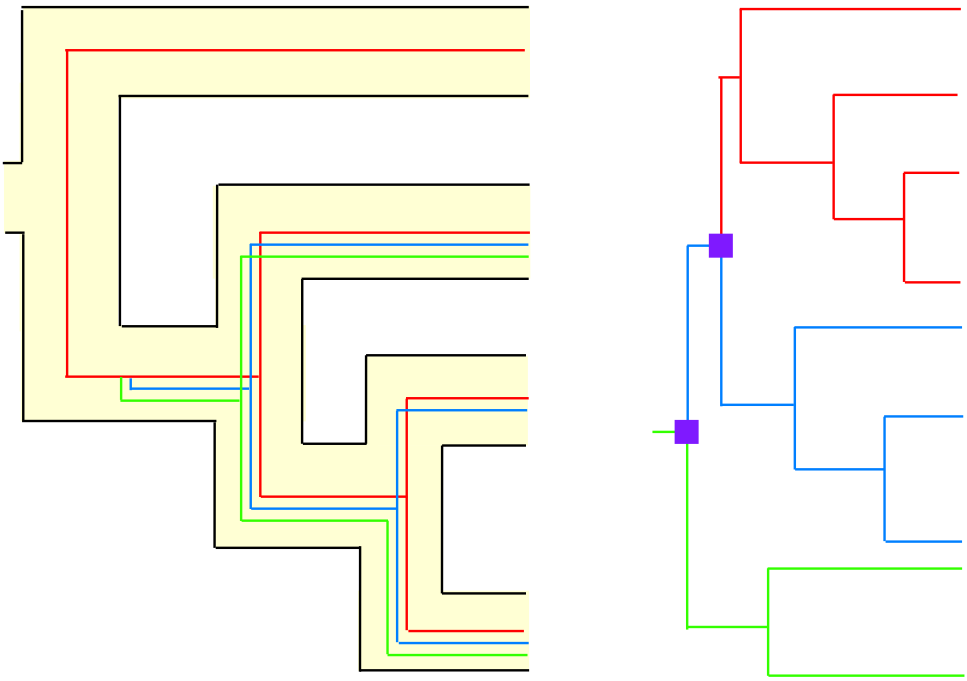

Figure 1.8: **Gene tree/species tree reconciliation.** On the left is a species tree with four taxa colored to distinguish orthologous genes related by speciation events. Duplication events connect the three gene trees within the species tree. On the right is a reconciled gene tree, showing the three different clades of orthologous proteins, connected by duplication events (represented by purple squares). Species A has one protein in this particular protein family, species B has three proteins, species C has two proteins, and species D has three proteins. The gene tree for these homologous proteins is shown on the right. Note that you need two duplication events to explain every repeat of the species tree, as shown in purple squares in the gene tree, where the repeated species trees (some with gene losses) are the red, green, and blue clades.

> *ters should not be naively assumed to be functional equivalents. The biological complexities force us to consider functional mutation in a more structured, probabilistic light.*

A more empirical justification for reconciliation methods was performed in a recent study [Chen *et al.*, 2007], in which pairwise comparison methods appeared to have higher sensitivity in ortholog identification, but the reconciliation methods appear to have higher specificity. It is unclear whether their method of evaluation (using latent class analysis) favors one type of method over another, and in particular whether higher sensitivity actually indicates additional false positives. But, overall, InParanoid and OrthoMCL appear to perform well. For application to phylogenomic functional prediction, it may be advisable to favor specificity over sensitivity. The incremental computational cost of tree reconciliation is trivial once a phylogenetic tree has been constructed. At the same time, tree reconciliation-based ortholog identification can be complicated by a number of confounding factors, such as the presence of duplicate sequences in sequence databases or multiple isoforms of the same gene (either within a population or due to splice variation), which incorrectly appear to be inparalogs. Another problem with reconciliation methods is that they do not help to identify paralogs when, after a speciation event and a subsequent duplication event, the original orthologous gene is lost and its duplicate maintained [Tatusov *et al.*, 1997].

Ortholog identification in prokaryotes is complicated by the high frequency of horizontal gene transfer (HGT) [Koonin *et al.*, 2001], which creates an evolutionary path that is not actually tree-like. The impact of HGT on phylogenomic analyses is an incorrect phylogeny, leading to false positive identification of duplication events, and resulting in poor functional predictions based on inaccurate reconciled trees. Some methods (e.g., OrthoMCL) explicitly require the sequences to be from eukaryotic

species to eliminate the difficulties caused by HGT, but this may be overly restrictive depending on the composition of the multi-gene family.

Despite their relative accuracy, MCMC methods, such as the model of Arvestad and colleagues [Arvestad *et al.*, 2003], have not been applied in reconciling gene trees and species trees as frequently as the more heuristic reconciliation methods because of their computational cost: the methods are difficult to apply on families with more than about 20 proteins. Nevertheless, these more rigorous methods are useful for higher accuracy in phylogenomic methods for small protein superfamilies.

1.5.5 Step 5: Overlay molecular function annotations onto the phylogeny

1.5.5.1 Motivation, definition, and methods

There are several publicly available general-purpose protein sequence databases with information on protein molecular function, including UniProt [UniprotConsortium, 2007], InterPro [Mulder *et al.*, 2007], and the Gene Ontology (GO) resource [Ashburner *et al.*, 2000]. The GO annotation (GOA) database [Camon *et al.*, 2004] provides molecular (biochemical) function, biological process and cellular localization annotations for specific protein sequences. Each annotation in the GOA database is accompanied by evidence codes specifying the origin of the annotation (e.g., *IEA* indicates inferred from electronic annotation, and *IDA* indicates inferred from direct assay) enabling biologists to distinguish between annotations based on experiment and those based on automated annotation methods. The SwissProt resource [Apweiler *et al.*, 2004] contains high-quality, manually-curated annotations for many sequences, but does not have a formal ontology. Specialized databases for model organisms are a good source for high-quality data for these species; examples include The Arabidopsis Information Resource (TAIR) [Garcia-Hernandez *et al.*, 2002], WormBase [Schwarz

et al., 2006], and FlyBase [Crosby *et al.*, 2007].

1.5.5.2 Limitations and considerations

The two main problems with existing annotation databases are the lack of experimental data and the existence of annotation errors. The GOA database is relied upon as a source of high-quality annotations, the GOA database reports that, as of October 13, 2007, less than 0.1% of the available annotations for UniProt proteins are derived from a method other than *IEA*. Estimates of existing error rates in annotation databases range from 8–40% [Brenner, 1999; Devos and Valencia, 2000], and experimental evidence is not immune from error. Because of the complete dependence of the phylogenomic protocol on accurate, experimentally-derived functional annotations as an annotation transfer method, this step is a primary source of possible inaccuracies.

1.5.6 Step 6: Infer function for the query protein

1.5.6.1 Motivation, definition, and methods

When phylogenomics is performed manually, the molecular functions are generally transferred within orthologous subtrees. Most automated programs mimic this methodology using a rule-based approach or a statistical method. The phylogenomic protocol would ideally take into consideration the relative density of annotations, how specific the functional terms are, and the evolutionary distance between the sequences when estimating the evolutionary persistence of the available annotations.

1.5.6.2 Limitations and considerations

Transferring function annotations within orthologous clusters is a risky endeavor because of the possible errors in identifying the clusters. Neither databases of orthologous proteins, such as COG [Tatusov *et al.*, 2000] or RoundUp [DeLuca *et al.*, 2006],

nor automated methods, such as Orthostrapper [Storm and Sonnhammer, 2002], take into account the non-transitive relationship of orthology. These methods often set the criterion for orthology low to account for this logical error, and thus the clusters inferred by automated methods may contain paralogs or missing orthologs.

The annotation transfer protocol must be defined appropriately. If a subtree has low bootstrap support, is it reasonable to transfer annotations within that subtree, even if all sequences in the subtree are related by speciation events? Is it reasonable to transfer all annotations within a particular cluster, regardless of the evolutionary distance spanned by that cluster? If a cluster contains a few recent duplication events (i.e., inparalogs), is it valid to transfer annotations within that cluster? If a probabilistic method is employed, will all annotations above a certain cutoff be transferred, or will only the annotation with the highest probability be transferred? All of these decisions depend on the evolutionary persistence of a particular molecular function, and whether the estimates of persistence are reasonable with respect to possible errors or inaccuracies in the reconstructed evolutionary relationships.

1.6 When is a phylogenomic analysis reasonable and effective?

Phylogenomic inference of protein molecular function is clearly only applicable under certain conditions: homologs to a query sequence of interest can be identified, reliable experimental functional annotations are available for at least some of these homologs, and the evolutionary distance between the query and homologs is not so great that the likelihood of functional divergence is high. In addition, standard approaches to phylogenomic inference depend on reconciling a species tree and gene tree. If the species of origin is not known (as in environmental sequence data [Yooseph *et al.*,

2007]), or if the species tree is ambiguous or inaccurate, it may be difficult to localize gene duplication events with sufficient accuracy.

Phylogenomic inference is complicated by parallel evolution and proteins with multiple functions. *Parallel evolution*, or the independent evolution of a single molecular function multiple times in a phylogeny, often complicates phylogenomic analysis. Enzyme substrate specificity in a subset of the aminotransferase family [Muratore *et al.*, 2007] (as described in detail in Chapter 3) is an example of this type of independent appearance of a function. Similarly, *moonlighting proteins*, or proteins that perform two or more different molecular functions under different conditions, and multifunction proteins may produce apparently inconsistent annotations within orthologous clades. These situations complicate, but do not preclude, phylogenomic analyses.

1.7 Evaluating phylogenetic tree reconstruction methods

The standard approach used to assess the accuracy of phylogenetic tree reconstruction methods is the use of simulation studies (e.g., [Huelsenbeck, 1995]). While these studies are useful, current simulation studies do not sufficiently evaluate the issues that arise in protein superfamily evolution such as structural divergence, lack of positional homology, and site- and lineage-specific rate variation. Since the accuracy of a phylogenetic tree topology is critical to phylogenomic analysis, and phylogenetic tree reconstruction methods do not always agree in critical aspects of the predicted tree topologies given the same input, some means of estimating the expected accuracy of these alternative methods would be helpful. For single-gene families, phylogenetic tree reconstruction methods are assessed in two ways: by way of simulation studies,

and by comparison of gene trees to species trees. However, in protein superfamily reconstruction the situation is more complicated. First, few simulation studies assess the effect of alignment errors on tree topology, or assess the ability to reconstruct phylogenies for sequences with diverged functions and structures. Direct analysis of predicted phylogenies for real superfamilies is similarly challenging when duplication events are included, since the actual evolutionary history of a protein family is unknowable.

Instead, we can use molecular function information to evaluate phylogeny reconstruction. For superfamilies, instead of using fossil or morphological character data to inform our reference tree topology, we can use experimental evidence of function and structure associated with the members of a protein superfamily. Under the assumption that evolution conserves function, a phylogenetic tree that is consistent with these experimental data is more likely to correspond to the true evolutionary history than one that is not. Of course, such analyses may sometimes be confounded by parallel evolution, horizontal gene transfer and ambiguous or incorrect experimental data. However, such comparisons of superfamily phylogenies with experimental data provide a direct measurement of the actual predictive power of phylogenetic tree reconstruction methods in application to phylogenomic inference. This performance measure is not without precedent; the OrthoMCL paper [Li *et al.*, 2003], for example, assesses their orthologous clusters based on how reliable the transferred EC numbers are within those clusters. Furthermore, there are a number of gold-standard datasets of protein superfamilies with experimentally validated functions (e.g., the Structure-Function Linkage Database [Pegg *et al.*, 2006]), that can be used for explicit phylogenomic evaluation. Such analyses would be complementary to more sophisticated simulation studies.

1.8 Thesis outline

Phylogenomic methods for protein molecular function prediction show great promise as a precise method to annotate protein molecular function when there is information on the molecular function of homologous proteins. As more protein sequences are added to databases and homology information improves, and more experimental evidence for molecular function is obtained, these methods will only improve in reliability and performance.

The rest of this thesis proceeds as follows. Chapter 2 introduces a method for phylogenomic analysis, SIFTER, and presents data regarding SIFTER's performance on a variety of protein families. Chapter 3 presents an additional method for use in conjunction with SIFTER, namely an active learning method using a mutual information criterion, that iteratively selects the protein function to experimentally characterize such that the amount of uncertainty regarding the functional predictions of the remaining proteins in the family is maximally reduced. Results of this active learner are also presented for a variety of protein families. Chapter 4 presents a study of SIFTER's performance on the annotation of 46 fully sequenced fungal genomes, including data for particular families and more general data illustrating SIFTER's overall performance. The final chapter sums up the contributions of this work.

Chapter 2

SIFTER: Statistical Inference of Function Through Evolutionary Relationships

2.1 Overview

The main work of this thesis is to describe the Statistical Inference of Function Through Evolutionary Relationships (SIFTER) method. SIFTER uses a statistical graphical model that applies principles from phylogenomics to automate precise protein function annotation [Engelhardt *et al.*, 2005; Engelhardt *et al.*, 2006]. We describe the most recent version of SIFTER (version 1.2), which is applicable to large and functionally diverse protein families because it includes a more general model of protein function evolution and a fast method for approximate calculation of posterior probabilities. We validated SIFTER on three diverse protein families: the AMP/adenosine deaminases, the sulfotransferases, and the Nudix family. SIFTER version 1.2 performed comparably to SIFTER version 1.1 when applied to the

AMP/adenosine deaminase family of proteins, with 93.9% accuracy on an experimental data set (where BLAST achieved 66.7%). On the functionally diverse sulfotransferase protein family, SIFTER achieved 70.0% accuracy (where BLAST achieved 50.0%) on experimental data. The sulfotransferase family also showed that the approximate computation of posterior probabilities works reliably across the full range of approximation granularities. On the exceptionally functionally diverse Nudix protein family, which was previously inaccessible to SIFTER because of the 66 possible molecular functions, SIFTER achieved 47.4% accuracy (where BLAST achieved 34.0%) on experimental data.

2.1.1 Protein function prediction

Automated protein function prediction is an exceptional challenge for computational biologists because protein function is difficult to describe and represent, protein databases are littered with annotation errors, and our understanding of how molecular functions arise and mutate over evolutionary time is far from complete.

The sequences of over 10^7 proteins are known, and a diverse array of functional descriptions have been attributed to these proteins, including 7466 molecular function terms from Gene Ontology [Ashburner *et al.*, 2000]. However, fewer than 0.2% of the annotations for UniProt proteins involved human curation in the Gene Ontology Annotation (GOA) database [Apweiler *et al.*, 2004; Camon *et al.*, 2004], and even fewer involved an experimental assay. Because biologists depend upon protein function annotations for insight and analysis, automated methods have been used to compensate for the relative dearth of experimental characterizations. Unfortunately these methods are commonly assessed based on annotation quantity rather than quality, resulting in a burgeoning of methods that increase the number of false positive function predictions. These results contaminate protein analyses and pollute

databases [Galperin and Koonin, 1998; Brenner, 1999].

A decade ago, when the protein sequence databases were small and mostly manually curated, Eugene Koonin estimated that the majority of errors in protein function annotation are actually propagations of existing database errors [Koonin *et al.*, 1996]. That is, the protein to be annotated has the same function as that of the matched database protein, but the protein in the database had been incorrectly described. This problem could be managed in part by having every protein annotation supported by traceable evidence. This would allow each protein annotation to be associated with a degree of confidence and would allow propagation of corrections to follow propagated errors. An important step in this direction is the GOA database, which incorporates the GO evidence codes and provides functional information for millions of proteins [Camon *et al.*, 2004]. The task of incorporating all literature evidence into the databases is immense and ongoing, but vital. Function prediction methods that incorporate evidence codes and provide reliability measures would seem less prone to error propagation.

Some automated methods have improved the quality of annotations by explicitly sacrificing functional specificity, making predictions at intermediate nodes of GO rather than at the leaves (e.g., GOtcha [Martin *et al.*, 2004]). These approaches are promising, though it remains to be seen whether the GO directed acyclic graph (DAG) is a satisfactory representation for generalizing molecular function and evolutionarily accessible functional variability. Our functional analysis of the Nudix protein family illustrates that the GO term coverage and hierarchical structure is incomplete and ineffective for some protein families.

2.1.2 Phylogenomics review

Phylogenomics has been proposed as a powerful approach for meeting the challenges of protein function prediction, as discussed in the introduction [Eisen, 1998; Brown and Sjolander, 2006]. The phylogenomic methods that we focus on in the remainder of this thesis use a full reconciled phylogenetic history of a protein family to make protein function predictions, rather than pairwise sequence comparisons as for predictions obtained from BLAST. This protocol relies on the observation that functional divergence often follows a gene duplication event, because protein redundancy will allow mutation events that otherwise would have been selected against. Duplication events are annotated at the internal nodes of a phylogenetic tree by reconciling inconsistencies between the gene tree and the associated species tree, which identifies the likely nodes in the gene tree of duplication events [Goodman *et al.*, 1979; Page, 1998].

2.1.2.1 Phylogenomics versus pairwise annotation transfer methods

The phylogenomic approach to protein function annotation has many advantages over pairwise annotation transfer methods, a few of which were discussed in the introduction. In general, a phylogeny suggests an evolutionarily-principled means of integrating functional evidence, and in particular ways of specifying how accurate each data point is believed to be with respect to the query protein. Since orthology is not a transitive relationship, organizing groups of proteins in orthologous groups based on this pairwise relationship does not guarantee that all of the members of the group will be related by speciation events as opposed to gene duplication events. Instead of pairwise comparisons, a tree is the natural structure to specify and explore protein homology and functional relationships.

2.1.2.2 Phylogenomic methods

While originally applied manually, phylogenetically-motivated protein function prediction has now been deployed in automated methods. One such method, Orthostrapper [Storm and Sonnhammer, 2002], uses bootstrapping to identify statistically-supported orthologous clusters of proteins, and transfers function annotations within each of these clusters. The statistically-supported clusters tend to encompass a subset of the sequences in a few large clusters, so often multiple annotations are transferred within each large cluster. Further, as there is little experimental evidence for protein functions, Orthostrapper makes relatively few annotations when restricted to experimental evidence, but those predictions it makes usually include the correct annotation. EnsEMBL now also uses a tree for more accurate function transfer among orthologs [Hubbard *et al.*, 2006].

The method presented in this thesis, SIFTER (Statistical Inference of Function Through Evolutionary Relationships), is also based on phylogenomic principles, which we formalize within a probabilistic framework [Engelhardt *et al.*, 2005; Engelhardt *et al.*, 2006; Engelhardt *et al.*, submitted]. SIFTER uses a statistical model of molecular function evolution to incorporate annotations throughout an evolutionary tree, making predictions supported by posterior probabilities for every protein. Phylogenomics is predicated on the explicit assumption that a phylogeny reconstructed from protein sequence represents also how molecular function evolved within those sequences. Thus, we fix the tree structure to the phylogeny reconstructed from sequence data and employ a conditional probability model describing molecular function evolution. This statistical graphical model of molecular function evolution enables access to a broad set of statistical tools for computation of posterior probabilities of the molecular functions and parameter estimation.

Predictions from statistical graphical models are generally quite robust, which this

problem requires. In particular, each protein family has sparse functional annotations and noise in both the annotations and the reconstructed phylogeny, so the selected model must be robust to the input data. The graphical model architecture is by nature flexible in terms of integrating various data types from different sources in a natural and coherent way. We currently represent molecular function in Gene Ontology terms, enabling some understanding of how the terms are related through the directed acyclic graph (DAG) structure organizing the terms hierarchically, although this is not a requirement of the approach. We rely on the evidence codes in the GOA database as an indication of the reliability the functional annotations.

The challenge of functionally diverse protein families

Protein families such as the Nudix family are a challenge for any molecular function prediction method because of the large number of proteins, the enormous diversity of molecular function in the family as a whole, and the sparsity of available experimental characterizations. Figure 2.1 summarizes protein family size and functional diversity in the Pfam protein family database [Bateman *et al.*, 2002]. In Pfam release 20.0, there are 8164 protein families, 519 of which have more than 1000 member proteins. We anticipate that these numbers will continue to grow, with a single recent project roughly doubling the number of known peptides [Yooseph *et al.*, 2007]. Although 5411 of the families have no experimental molecular function characterizations from the GOA database, there are 1887 families with at least two different molecular functions based on experimental evidence. SIFTER could be applied to each of these families, producing predictions for more than one candidate molecular function. Of those families, 619 families (32.8%) have six or more different molecular functions characterized within the family's proteins. The SIFTER model nominally has exponential computational complexity in the number of candidate proteins, so

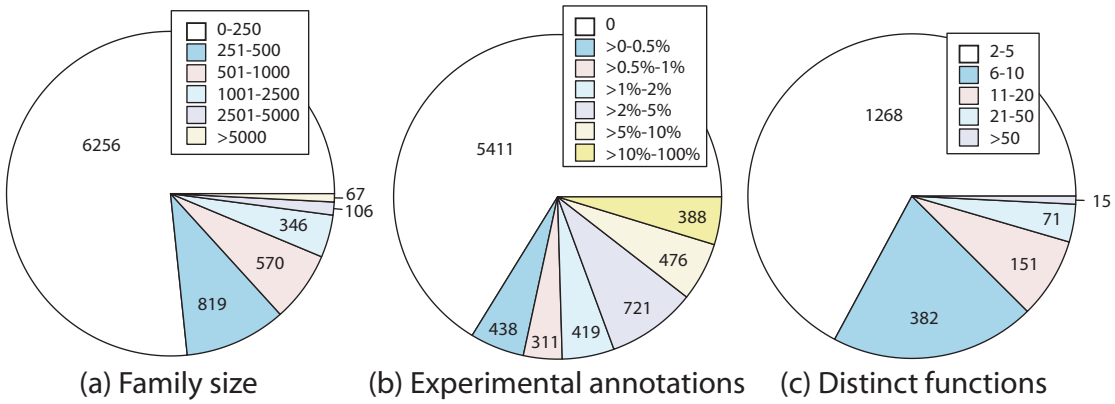

Figure 2.1: **Numerous and diverse protein families in the Pfam Database.** Statistics from Pfam release 20.0 show large and functionally diverse protein families, motivating an approximate version of SIFTER. Panel (a) shows the proportion of the 8164 protein families in Pfam that have the specified number of protein members, where 67 families have over 5000 members. Panel (b) represents the proportion of proteins with experimental annotations from the GOA database in each family within Pfam, including 5411 families with zero experimental annotations, and 388 families with more than 10% of their members with experimental characterizations. Panel (c) illustrates the functional diversity of the 1887 families with two or more different experimental function characterizations. Note that almost a third of those families have six or more functions, and would be computationally infeasible for SIFTER version 1.1.

most of these functionally diverse families are computationally infeasible for the previous version of SIFTER. This motivated the construction of a fast approximation of SIFTER to analyze these functionally diverse families with phylogenomic methods for more complete coverage of Pfam protein families. We predict that as the number of proteins with experimental characterizations increases, so will the relative diversity of the families (despite the experimental bias towards characterizing proteins with the same molecular function as evolutionarily close, characterized proteins).

This most recent version of SIFTER uses a simple but effective approximation that enables tractable computation of function predictions in functionally diverse families. The approximation essentially truncates the number of molecular function combinations that are considered during computation of the posterior probabilities. Despite

the simplicity of the approach, prediction quality does not appear to degrade until the lowest truncation is reached (and then only slightly), even for multifunction proteins. We validated this approach on the previously studied AMP/adenosine deaminase family of proteins, and then we applied it to two functionally diverse protein families, the sulfotransferase family and the Nudix family, the latter of which was previously computationally infeasible for SIFTER.

In this version of SIFTER, we designed the model of molecular function evolution to be flexible enough to enable the encoding of prior biological knowledge and to allow us to construct the transition rate matrix in a semantically meaningful way from a smaller set of parameters that can be estimated from the data. The model for evolution that fulfills these requirements is a generic continuous-time Markov chain. This also simplifies the machinery required to compute posterior probabilities and to estimate the model parameters.

2.2 The SIFTER method

SIFTER incorporates data from many different sources to reconstruct a phylogeny and compute posterior probabilities. Here we describe the data integration, then we present the Markov chain model and the approximate computation of posterior probabilities.

2.2.1 From database data to a tree

The data used by SIFTER currently comes from a number of different sources. We extracted the families studied here from the Pfam database [Bateman *et al.*, 2002], and we used the manually-curated alignment found in Pfam for phylogeny reconstruction. Trees were built using different methods depending on the family size as

Evidence Code	Full Name	Probability
IDA	Inferred by direct assay	0.9
TAS	Traceable author statement	0.9
IMP	Inferred by mutant phenotype	0.8
IGI	Inferred from genetic interaction	0.8
IPI	Inferred from physical interaction	0.8
ISS	Inferred by sequence or structural similarity	0.4
RCA	Inferred from reviewed computational analysis	0.4
IGC	Inferred from genetic interaction	0.4
IEP	Inferred from expression pattern	0.4
IC	Inferred by curator	0.4
NR	Not recorded	0.3
NAS	Non-traceable author statement	0.3
ND	No biological data available	0.3
IEA	Inferred by electronic annotation	0.2

Table 2.1: The Gene Ontology evidence codes and corresponding SIFTER probabilities of correctness. These probabilities were elicited from a domain expert.

described below. All trees were reconciled using the Forester v.1.92 program [Zmasek and Eddy, 2001a]; the reference species tree is from the Pfam database. As input evidence, we used all annotations in the GOA database [Camon *et al.*, 2004] with experimental evidence codes *IDA*, *IMP*, and *TAS*. Where we independently found experimental characterizations in the literature, we labeled that annotation with a Traceable Author Statement (*TAS*) evidence code. The probability of correctness for each of the evidence codes is shown in Table 2.1. These probabilities were elicited from a domain expert (Professor Steven Brenner).

In the model, each protein i is associated with a Boolean random vector X_i, where each Boolean random variable represents a candidate function that takes value 1 when protein i has that particular molecular function and 0 if that function is not active in protein i. The candidate terms and associated annotations from the GOA database are identified and converted to the random vectors X_i in the tree associated with each

protein i through the following process. We eliminate the molecular function term dependencies and reduce the number of candidate functions by annotating the GO DAG terms with experimental evidence codes for the entire family of proteins, and first pruning all (possibly annotated) ancestors of annotated nodes, then pruning all non-annotated nodes. This leaves a set of *candidate functions* that are neither ancestors nor descendants of each other, ensuring there are no deterministic dependencies between them in terms of the semantic network. Then, for each protein with experimental evidence, the annotations from the full GO DAG are propagated to the set of descendant candidate functions by effectively marginalizing out the ancestor terms. Annotations are propagated to the candidate terms by assuming that the probability that children terms have a value 1, when a parent term has value 1 and the edges between related terms are all "is a" edges, is $\frac{1}{r^{|S|}}$. In this equation, $|S|$ is the size of an arbitrary subset of children terms of the annotated term and r is the solution to the equation $\sum_{S \in \mathbf{S}} \frac{1}{r^{|S|}} = 1$, where $\mathbf{S}$ is the power set of all children terms of a particular term. Note that we set the probability of the empty set to zero, effectively assuming that if a protein has a particular function, it must also have at least one of the function's descendant terms related by "is a" edges. Marginalizing out all of the non-candidate function terms eliminates all deterministic dependencies from the random vector for each protein. The random vectors representing observations of molecular function activity are set to the values from this computation for each leaf protein with experimental evidence.

We then propagate the evidence throughout the phylogenetic tree to compute posterior probabilities for all of the proteins in the tree. We do this by using a Markov chain model representing how protein molecular function evolves to define transition probabilities associated with the branches of the tree, and by applying standard message passing techniques (e.g., [Felsenstein, 1989]) to compute posterior probabilities at all nodes in the tree

2.2.2 Markov chain model

The structure of the evolutionary model is given by the phylogenetic tree; we are left to specify the conditional probabilities at each of the nodes in the tree. SIFTER employs a first-order Markov chain, which makes the Markov assumption that future states are independent of past states given complete information about the present state. In evolutionary terms, this means that predicting molecular function for a protein based on only its immediate ancestor will be just as good as using all of its ancestors. Similar assumptions have long been used to model the evolution of molecular sequences (e.g., [Dayhoff *et al.*, 1978; Henikoff and Henikoff, 1992]). Let the number of candidate functions in the model be denoted M, and the number of leaf proteins be denoted N. Our Markov chain has three sets of parameters. Two parameters $\sigma_{speciation}$ and $\sigma_{duplication}$, as in earlier models, describe the scaling of time after speciation versus duplication events; vector parameter $\alpha = \{\alpha_1, ..., \alpha_M\}$ describes the rate for a single instantaneous time step of function i arising when none of the candidate functions are observed in the ancestor protein; in matrix parameter $\Phi = \{\phi_{11}, \phi_{12}, ..., \phi_{MM}\}$, the off-diagonal elements $\phi_{ij}, i \neq j$, describe the rate for function i mutating to also perform function j, and the diagonal elements ϕ_{ii} describe the rate at which function i is lost.

Using parameters Φ and α, we can build the instantaneous transition rate matrix Q for a Markov chain that describes the instantaneous rate of change in functional activity from an ancestor to its descendant. For example, the matrix Q for two candidate functions has on its rows (representing ancestor states) and columns (representing descendant states) all of the possible combinations of two functions for the parent and child proteins, namely $\{00, 01, 10, 11\}$. Call this power set S_2, as above. In this terminology, "01" (for example) means that the first candidate function is not present (0) and the second candidate function is present (1) in a protein's state.

	00	10	01	11
00	–	α_1	α_2	0
10	ϕ_{11}	–	0	$\phi_{12} + \alpha_2$
01	ϕ_{22}	0	–	$\phi_{21} + \alpha_1$
11	0	ϕ_{22}	ϕ_{11}	–

Table 2.2: SIFTER **rate transition matrix.** This is the instantaneous transition rate matrix used in SIFTER for $M = 2$. The rows represent different sets of functions for the parent protein; the columns are child functions. Recall that α_i is the rate of a function i arising, ϕ_{ij} is the rate of a protein with function i mutating to also have function j, and ϕ_{ii} is the rate of function i disappearing. The diagonal elements in this table are set such that the rows sum to 0.

A full transition rate matrix Q, defined by variables Φ and α, is shown in Table 2.2 for $M = 2$, and is constructed analogously for $M > 2$ candidate functions. We will motivate and discuss this particular transition rate matrix further below. For a thorough discussion of continuous-time Markov chains as related to evolutionary processes, see [Felsenstein, 2003], Chapter 13.

We constrain the rows of the matrix Φ to positive values that sum to at most M. Similarly, we constrain the α parameters to the M simplex. These constraints must be respected by the parameter estimation procedure. Furthermore, we constrain these parameters away from zero so as to avoid creating sink states or unreachable states in the Markov chain when estimating parameters (specifically, all parameters are greater than 0.01). Because the Φ matrix and vector α are used to construct the transition rate matrix Q, each non-zero, off-diagonal entry in matrix Q implicitly has a finite upper bound and positive lower bound by way of the constraints on the matrix Φ and the vector α.

The conditional probability of a child configuration given a parent configuration is as follows, from the definition of a continuous time Markov chain:

$$p(X_i = s_j | X_{\pi_i} = s_k, t_i, \Phi, \sigma) = \{\exp(t_i \sigma Q)\}_{k,j}.$$

In this equation, X_i is the random vector associated with protein i, X_{π_i} is the random vector associated with protein i's immediate ancestor, t_i is the length of the branch between X_i and X_{π_i}, j and k index the power set $\mathbf{S}$, s_j and $s_k \in \mathbf{S}$, and exp is the matrix exponential function, $\exp(tQ) = \sum_{k=0}^{\infty} \frac{(tQ)^k}{k!}$. Element (k, j) of the transition probability matrix $\exp(t_i \sigma Q)$ is the probability that an ancestor protein in state k mutates to state j in the descendant protein within the time period t_i (here, the phylogenetic branch length, which must be non-negative).

As in earlier models [Engelhardt *et al.*, 2005; Engelhardt *et al.*, 2006], the joint probability of the complete tree is:

$$p(X|\Phi, \sigma) = p(X_{root}) \prod_{i \in tree} p(X_i | X_{\pi_i}, t_i, \Phi, \sigma).$$

The parameters of this model can be estimated using generalized expectation maximization (GEM) [Gelman *et al.*, 2003]. In particular, the E-step is the computation of the posterior probabilities for each random variable, using the standard message passing algorithm as in Felsenstein's maximum likelihood tree estimation [Felsenstein, 1989]. Because there is no simple analytical expression for the matrix exponential function of this transition rate matrix Q, we compute these values numerically for a given Q using the jLapack library [Blount and Chatterjee, 1998].

The M-step is implemented using projected gradient ascent [Bertsekas, 1999] for each of the parameters σ, Φ, and α, derived from the gradient of the expected complete log likelihood of the model with respect to each of the parameters. Each step of the gradient ascent is scaled by step size ρ. The parameter constraints mentioned above define the space onto which the gradient steps are projected. The Φ and α parameters are projected via normalization onto an $M + 1$ sided cone defined by the M simplex

on M dimensions, and the scale parameters are projected back to 0.01 when they fall below that value. Because the parameter gradients are unbounded objective functions in this optimization, the constraints are necessary to bound these functions during parameter estimation.

In practice, we take a single projected gradient step for each iteration of GEM. We stop EM iterations when the sum of the absolute value of the total change in parameters is less than some cutoff c. In our experiments here, we set the step size ρ of the gradient ascent to 0.01 and the cutoff c to 0.0015 but these will vary based on the size of the family and the number of observations.

2.2.3 Transition rate matrix assumptions and alternatives

We designed the instantaneous transition rate matrix Q with the following semantics. In a single instant, the probability of more than one functional change (i.e., loss or gain of a single function) in a protein is zero. Hence those instantaneous transition rates are set to zero. Of course, the probability of these transitions will be non-zero when time $t > 0$ has passed, according to the definition of the matrix exponential. Another way to see this is that the probability of multiple transitions (creating a path between the states with more than one functional change) will be non-zero when some finite period of time has passed. Furthermore, some states are the result of one of multiple possible events. For example, if a parent protein in state 01 transitions to state 11 in the child, the appearance of the function 1 could be a result of function 2 mutating into function 1 while retaining function 2 as well (ϕ_{21}) or the spontaneous appearance of function 1 (α_1). For transitions that could result from one of a number of events, the rates for each possible event are summed.

This approach thus takes into account the possibility of a single change in function over a finite time period. This models the impact of various changes in protein

sequence that control and modify function. An additional domain may be added to a protein in a single mutation event (i.e., a gene duplication or domain shuffling event), conferring an additional molecular function. Mutations of individual nucleic acids (coding for this protein or related proteins) or a change in environment may accumulate to confer enzymatic activity specific to an additional substrate, or yield (over time) a different chemical reaction entirely. All of these possibilities are implicitly modeled by our particular choice of matrix Q.

Other evolutionary possibilities are not modeled by our choice of matrix Q. In particular, we have assumed that the instantaneous rate of transition between states with more than one difference, e.g., a 01 state and a 10 state, has probability zero. Of course this does not reflect all biological possibilities. There are examples of single nucleotide mutations, an event that would be considered instantaneous, that change specificity from one substrate to another. We have chosen to allow this case to be subsumed by the transition paths implemented by the matrix exponential, in particular a function gain followed by a complementary function loss.

A more general modeling concern may be the simplification of describing a protein performing a certain function as a binary variable. Alternatively, we could model this using a continuous variable capturing the tendency of a particular enzyme to catalyze a particular reaction, such as K_{cat}. It would be possible to use diffusion theory to model this variable as a continuous one, but we have chosen not to go this route for a number of reasons. The primary reason is one of data: there is simply not enough data available for particular enzymes to model this robustly. A more subtle question is whether this feature of a protein evolves in parallel with protein sequence, which impacts the appropriateness of phylogenetic methods for this modified problem.

2.2.4 SIFTER's approximate computation

The time complexity of computing the posterior probabilities for the SIFTER model is linear in the number of proteins, but exponential in the number of candidate functions. The exponential complexity is due to the set of 2^M possible combinations of candidate functions for each protein, resulting in a transition rate matrix with $(2^M)^2$ entries. Computing matrix exponentials explicitly for each branch in the tree (after storing intermediate results) has a computational complexity of $O(N((2^M)^{2\cdot 3})$ (the 3 comes from computing the matrix exponential of a matrix of size $(2^M)^2$). Thus, when M is large, the time to compute posterior probabilities is dominated by the exponential complexity in the number of candidate functions. In the case of the Nudix family, where the number of candidate functions is 66, this complexity would make the computation impossible. But most of the configurations under consideration, especially states with many functions having value 1, have a near-zero probability of occurring. Therefore, we implemented a simple but effective approximation that truncates the possible sets of molecular functions under consideration.

The power set truncation approximation simply limits the total number of candidate functions with value 1 in the transition rate matrix power set to a value T. This shrinks the transition rate matrix to only consider transitions between all states with T or fewer functions with value 1. In the binary representation above, the sum of the candidate functions with value 1 (i.e., present in the protein) cannot exceed a fixed value T, thus reducing the number of elements in the power set to $\sum_{i=1}^{T} \binom{M}{i}$. In the Nudix family, setting $T = 1$, the power set has 67 elements (each single candidate term and the empty set), and 4489 elements in the associated transition rate matrix Q, where the power set truncated at 2 has 4356 elements, and a Q matrix with $18,974,736$ elements. In contrast, without truncation, the Nudix power set has approximately $7.38 * 10^{19}$ elements, and the corresponding Q matrix has approximately

$5.44 * 10^{39}$ elements. Computation time is reduced from infeasibly long to seconds, as illustrated in the results section.

2.3 Results and discussion

We first compare the performance of SIFTER (version 1.2) on a previously studied family (the AMP/adenosine deaminases) to the performance of SIFTER version 1.1 [Engelhardt *et al.*, 2006]. We then present results on the sulfotransferase and Nudix families to assess whether the model is able to predict molecular function accurately for more functionally diverse protein families.

In these analyses, we define SIFTER *accuracy* as the percentage of experimentally characterized proteins for which the functional term with the maximum posterior probability is in the set of experimental characterizations. All experiments where timing results were reported were run on Dell Precision 390 Workstation computers with Intel Core2Duo 2.6 GHZ processors and 2 Gb RAM.

2.3.1 Comparative function annotation methods

Before presenting the results, we explain how we ran each of the function annotation methods used for comparison. In general we ran these methods giving them the benefit of the doubt, choosing parameter settings to to make the comparison as fair as possible. Of the three methods involved in the comparison, BLAST and GOtcha both transfer function annotations based on pairwise sequence comparisons; Orthostrapper comes from a family of methods that rely on phylogenomic assumptions, transferring annotations between proteins that have pairwise orthology with bootstrap significance.

The BLAST version 2.2.4 [Altschul *et al.*, 1990] assessment was performed on the

non-redundant (nr) set of proteins downloaded from the NCBI website on December 11, 2006. We ran BLASTP with an E-value cutoff of 0.01. For each query protein in the selected families we searched the BLAST output with the most significant E-value (probability of the alignment score based on an extreme value distribution for aligning protein sequences at random) removing any exact matches from the same species to ensure that the query protein did not receive its own database annotation (emulating a leave-one-out type of experiment). We transferred the candidate functional term associated with the most significant BLAST hit that had an annotation within the set of candidate functions defined by SIFTER for a particular protein family. In other words, if the top BLAST hit was not annotated, we found the next most significant hit with a candidate function annotation to transfer. If there were multiple proteins annotated with candidate functions that shared the same E-value, we transferred all of the associated annotations. Historically, when researchers use BLAST for large-scale protein molecular function annotation, either the most significant non-identity hit or the most significant non-identity annotated hit is transferred to the query protein, often leading to no prediction or an incorrect (or overly general) prediction. Here we transfer annotation from the most significant hit with a candidate term using a keyword extraction script. This process increases the overall accuracy of the BLAST predictions and enables a comparative ROC-type analysis. This problem requires a ROC-type analysis that allows for the possibility of multiple true positive functional predictions and multiple true negative functional predictions, instead of the usual binary classification. Specifically, the ROC-type analysis was performed by determining true positive and false positive annotations for E-value cutoffs between 0 and 0.01.

We ran the first publicly available version of the GOtcha software [Martin *et al.*, 2004]. GOtcha predicts protein function using a statistical algorithm applied to BLAST searches. The BLAST searches are performed on a manually-constructed

database containing complete GO annotations of seven genomes, including GO evidence codes. Because the annotation database is precompiled for fast querying, we could not ensure that a query protein was not being annotated from its own annotation in the database; thus our results for GOtcha are likely to be overly optimistic. For one set of experiments (labeled GOtcha), we made predictions using annotations with both experimental and electronic evidence codes. For another set of experiments (labeled GOtcha-exp), we made predictions given only annotations with experimental evidence codes. The output is a numerically ranked list of GO terms; we extracted the ranked list of candidate functions from this complete set, breaking ties in favor of the correct term. The ROC-type analysis was performed by determining true positive and false positive annotations for cutoff values between 100 and 0 (maximum and minimum ranking scores, respectively).

We ran the Orthostrapper [Storm and Sonnhammer, 2002] version from February 6, 2002. We split the proteins in each family with experimental GO annotations into proteins from eukaryotes and non-eukaryotes, which was used to determine orthology or paralogy using an approximate gene tree/species tree reconciliation method. We clustered the bootstrap analysis according to the cluster program in Orthostrapper, using a bootstrap cutoff of 750 and then using a cutoff of 1, resulting in statistically significant clusters (Orthostrapper-750) and non-statistically significant clusters (Orthostrapper-1). In each cluster, we transferred all experimental GO annotations from member proteins onto the remaining proteins without experimental characterizations. If a protein was present in multiple clusters, it received annotations transferred within all of those clusters. This method yields an unranked set of predictions for each protein; multiple annotations were manually resolved in favor of the correct one. We performed cross-validation for each protein by removing its annotations and transferring the remaining annotations to make a prediction for the held-out protein. The ROC-type analysis was performed by determining true positive and false positive

annotations for all clusters generated by bootstrap cutoffs between 1000 and 0.

2.3.2 AMP/adenosine deaminase family

We applied SIFTER version 1.2 to the Pfam adenosine/AMP deaminase family (PF00962), which contains 251 proteins in Pfam 18.0. Proteins in this family are responsible for removing an amine group from the purine base of three possible substrates: adenine, adenosine, and AMP. As shown in Figure 2.2, there are four candidate functions for this family, three of which are deaminase activity with different substrates. Additionally, a subset of proteins within this family show growth factor activity, and are commonly known as adenosine deaminase-related growth factors [Maier *et al.*, 2005].

The GOA database contained experimental GO annotations for 13 proteins, a literature search revealed experimental annotations for an additional 20 proteins, including our experimental characterization of an adenosine deaminase protein in *Plasmodium falciparum* [Engelhardt *et al.*, 2005], resulting in 33 proteins with experimental annotations. A phylogeny for these 33 proteins, in Figure 2.2, shares the branching structure with the reconstructed phylogeny in a previous study [Maier *et al.*, 2005] regarding the relative positions of the adenosine deaminases, adenine deaminases, AMP deaminases, and adenosine deaminase-related growth factors. Both this phylogeny and the phylogeny for the full set of proteins was built using PAUP* version 4.0*b*10 maximum parsimony with the BLOSUM50 matrix [Swofford, 2001; Henikoff and Henikoff, 1992]. The alignments for both phylogenies were from the Pfam alignment.

Besides being an important protein family in the study of human immunodeficiency disease [Hirschhorn and Ellenbogen, 1986], this family is interesting in the molecular function context because the active site residues are shared across all of the different types of substrates (i.e., in all cases the site binds to an amine) [Ribard

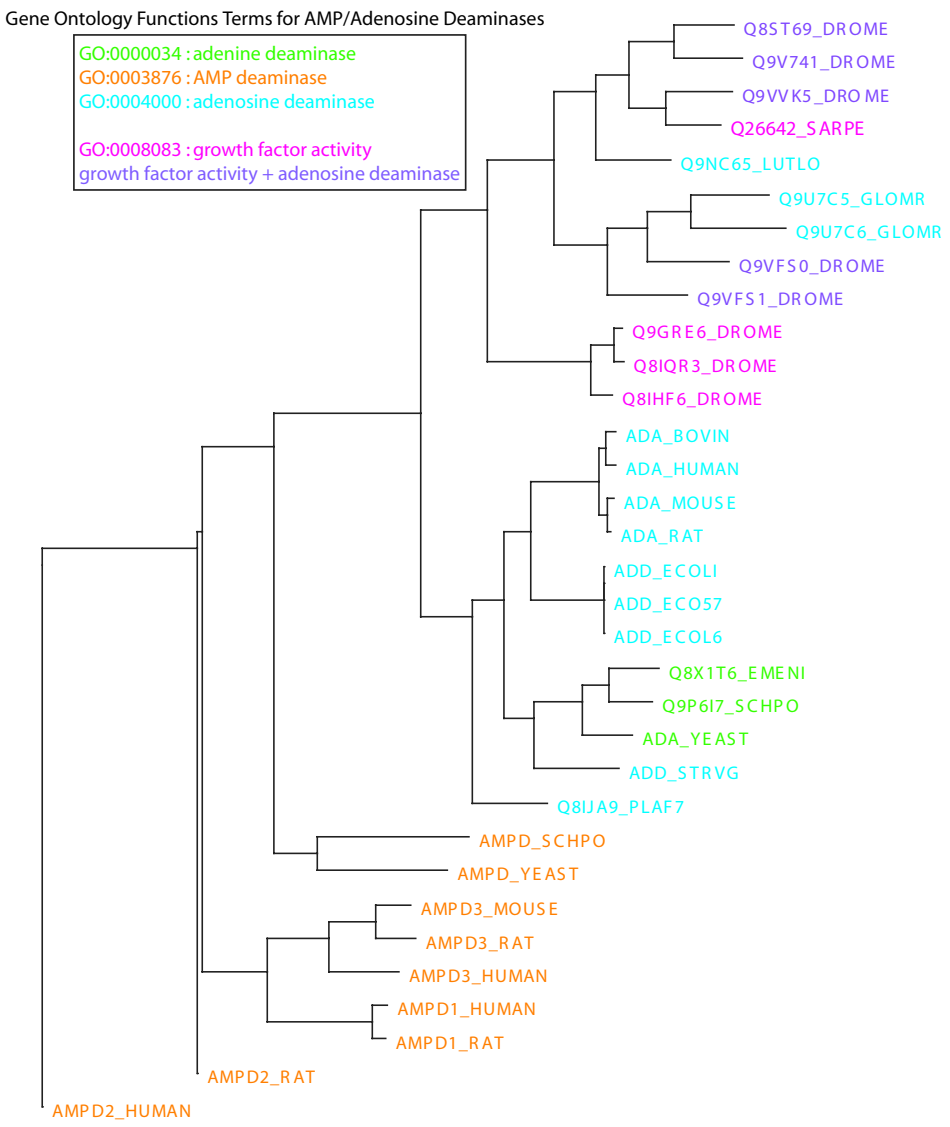

Figure 2.2: **Phylogeny of experimentally characterized AMP/adenosine deaminase proteins.** The phylogeny of the experimentally characterized set of proteins from the AMP/adenosine deaminase family. We built the phylogeny by extracting the protein sequences for the experimentally annotated proteins from Swiss-Prot/TrEMBL, aligning them to the A_deaminase HMM profile using `hmmalign` [Eddy, 1998], and reconstructing a phylogeny using PAUP* parsimony. The branching structure is the same as that of the full tree used in the experiments at the top levels of the phylogeny. The colors indicate experimentally characterized protein functions, as described explicitly in the key. This tree was drawn using the ATV program [Zmasek and Eddy, 2001b] (as were all of the remaining trees presented here).

et al., 2003]; substrate specificity in this protein is modified by molecular changes in areas not associated with amine binding. Thus a closer look at the active site will not result in better discrimination of the protein substrate, only a general evolutionary divergence. A similar situation is also observed in the protein structures of the Nudix hydrolases.

We ran leave-one-out cross-validation on the experimental annotations, using the full phylogeny of 251 proteins in this family. In general, cross-validation assesses the ability of the method to generalize to unobserved data. Each iteration of cross-validation involved estimating the model parameters using GEM after removing evidence associated with one of the proteins in the model, simulating unobserved data. We call the prediction correct when the function with the maximum posterior probability for the held-out protein, computed using the estimated parameters, is one of its experimentally characterized functions. The starting values of the parameters were $\phi_{ii} = 0.5$, $\phi_{ij} = 1.0$ for $i \neq j$, $\sigma_{speciation} = 0.5$, $\sigma_{duplication} = 0.8$, and $\alpha_i = 1.0$.

Leave-one-out cross-validation on the experimental annotations yields 93.9% accuracy (31 out of 33 proteins with one of the associated experimental annotations having the maximum posterior probability). Of the two proteins with incorrect SIFTER predictions, one protein (Q9NC65_LUTLO) with adenosine deaminase activity located near the growth factor activity clade is incorrectly predicted to have growth factor activity [Charlab *et al.*, 2000], and one protein (ADD_STRVG) with adenosine deaminase activity is incorrectly predicted to have adenine activity. It is hypothesized that adenosine deaminase activity confers growth factor activity through the destruction of adenosine, which induces apoptosis in some types of cells [Maier *et al.*, 2001], so the experimental annotations for the proteins with only growth factor activity annotations may be incomplete.

2.3.2.1 Comparison of annotation methods

We ran BLAST, GOtcha, and Orthostrapper on the set of experimentally characterized AMP/adenosine deaminase proteins. The comparison on the experimental annotations show that BLAST and GOtcha-exp (i.e., GOtcha transferring only experimental annotations) achieve 66.7% accuracy (22 of 33), GOtcha (using all annotations) achieves 87.9% accuracy (29 of 33), and Orthostrapper-1 [Storm and Sonnhammer, 2002] (bootstrap cutoff of 1, meaning only one out of 1000 bootstrap trees must label two proteins as orthologs to put them in an orthologous cluster) achieves 78.8% accuracy (26 of 33). For GOtcha-exp, we broke ties in favor of the correct function 14 times over the 33 proteins.

The ROC-type analysis in Figure 2.3 is a better method for comparison in this multifunction family. Our measure of accuracy only takes into account the molecular function prediction with the highest posterior probability, and does not account for other highly ranked molecular functions. Instead, the ROC-type analysis evaluates true positives as compared with the percentage of false positives across all cutoff values, where a positive prediction is one that has a posterior probability above the cutoff. This explicitly accounts for all of a protein's experimentally characterized functions in a multifunction protein family when measuring performance. In the figure, SIFTER outperforms all of the methods on this family at all error rates except for a small section where Orthostrapper-1 has slightly better performance. Within the area of high specificity, which is often the most relevant area for quantifying performance on biological sequence analysis, SIFTER's performance advantage is striking.

2.3.2.2 Parameter estimation

We ran GEM to estimate the parameters for the AMP/adenosine deaminase family, including all of the available experimental annotations. Examination of the parame-

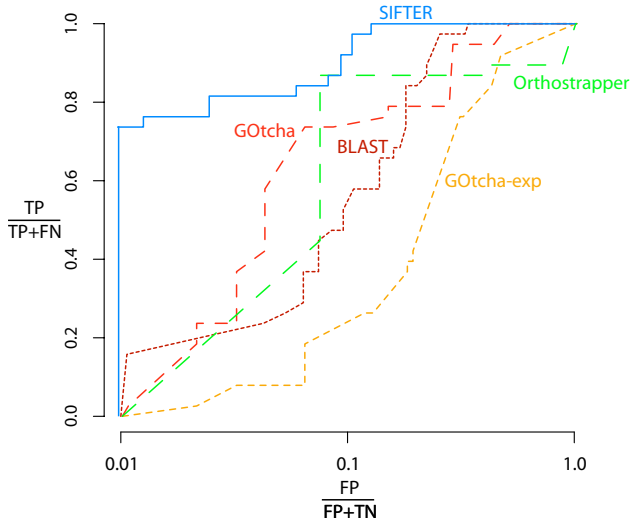

Figure 2.3: **Functional annotation methods comparison on AMP/adenosine deaminase family.** A comparison of results for SIFTER and other annotations methods on the AMP/adenosine deaminase family of proteins. This ROC-type analysis shows the rate of false positives versus true positives as the acceptance cutoff varies from admitting no annotations to admitting all annotations. SIFTER performs well relative to the pairwise annotation methods under this criterion. Note that the x-axis is on a log scale.

ter estimates for this family gives no obvious insight into how the functions evolved, and one should be wary of interpreting these estimated parameters in an evolutionary light. In particular, the parameter governing the spontaneous appearance of growth factor activity is estimated to be over 4 times lower than the corresponding parameter for the other three functions (0.288 versus 1.233 for adenine, 1.204 for AMP, and 1.275 for adenosine). It appears that the growth factors share a novel sequence motif, where two of the four conserved residues are also found in the adenosine and adenine deaminase proteins [Maier *et al.*, 2005]. This does not differentiate the evolutionary appearance of growth factor activity from substrate evolution in this family. It is possible that the parameter estimates imply that growth factor activity should not be modeled as arising spontaneously, but instead be modeled as evolving from a particular deaminase activity (in this family, adenosine). The scale factors $\sigma_{speciation}$ and $\sigma_{duplication}$ did not provide any evolutionary insight, as they both converged quickly to the boundary 0.01. On one hand, this suggests that the central role of gene duplication in phylogenomics may be overemphasized relative to the evolutionary history of actual function mutations; on the other hand, the large number of false positive gene duplication events in the reconciled trees produced through automated pipelines appears to mitigate or altogether eliminate their signal.

2.3.2.3 Comparison with previous SIFTER

We compared the version of SIFTER presented here(version 1.2) with the previous version of SIFTER (version 1.1) [Engelhardt *et al.*, 2006] on the AMP/adenosine deaminase protein family. We computed the accuracy for leave-one-out cross-validation on the deaminase protein family (running EM for each iteration, with no truncation), where SIFTER version 1.1 had 93.9% accuracy (31 of 33) and SIFTER version 1.2 had 93.9% accuracy (31 of 33), missing the same two proteins. The performance of the two methods are almost identical and show no relevant differences in the ROC-type

analysis (Figure 2.4).

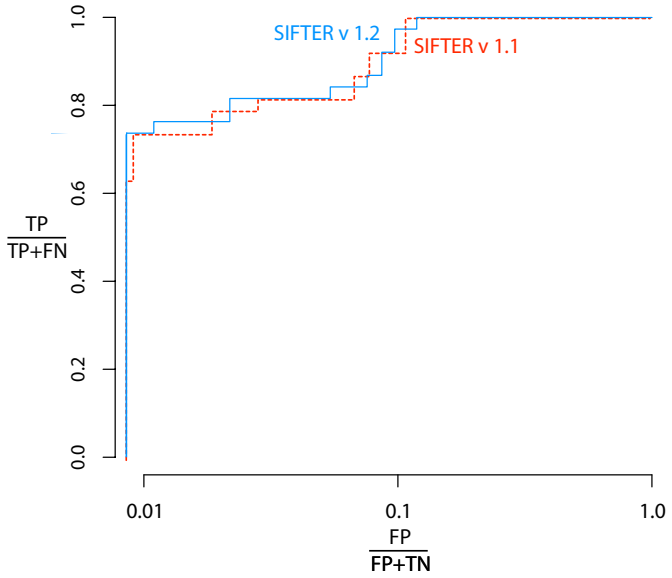

Figure 2.4: **ROC-type comparison of SIFTER version 1.1 and SIFTER version 1.2 on AMP/adenosine deaminase family.** A comparison of SIFTER version 1.2 with SIFTER version 1.1 on the AMP/adenosine deaminase family of proteins. The ROC-type analysis shows the rate of false positives versus true positives as the acceptance cutoff varies from admitting no annotations to admitting all annotations. The ROC-type curve for SIFTER version 1.1, as described in [Engelhardt *et al.*, 2006], is almost identical to that of SIFTER version 1.2, as described here. Note that the x-axis is on a log scale.

In terms of computation speed, SIFTER version 1.1 averaged 296.2ms with 41.6ms standard deviation for 10 iterations of exact computation on the deaminase family, whereas SIFTER version 1.2 averaged 455.3ms with 55.3ms standard deviation for identical 10 runs on the same computer. The maximization step for EM averaged 11.4ms for SIFTER version 1.1, and 13.8ms for SIFTER version 1.2. The comparative results for this particular family show that the performance of the two models are of similar magnitude.

2.3.2.4 Power set truncation approximation results

We used the AMP/adenosine deaminase family to test the power set truncation approximation. We computed posterior probabilities based on the parameters previously estimated with no truncation from the complete experimental data set, truncating the number of possible functions associated with a single protein at 1, 2 and 3. Figure 2.5a shows the number of predictions at the leaf proteins that differed (regardless of correctness) from the algorithm with no truncation (i.e., truncation level 4), for each of the three possible levels of truncation. Figure 2.5b shows the mean difference and variance in posterior probabilities for the leaf proteins at each level of truncation, as compared to the posterior probabilities computed without truncation at the leaf proteins. Figure 2.5c shows the average running time for all of the four possible levels of truncation, with the number of rows and columns of the transition rate matrix embedded in the bars. The impact on the posterior probabilities and corresponding functional predictions for a fixed set of parameters at all but level 1 appears modest.

An alternative test of the truncation approximation is to run leave-one-out cross-validation, estimating the parameters with the truncated algorithm at each iteration, for each of the truncation levels. Truncation levels 4, 3 and 2 all achieved 93.9% accuracy (31 of 33), whereas truncation level 1 achieved 90.9% accuracy (30 of 33), missing the additional prediction for protein Q26642_SARPE (predicting adenosine deaminase activity when it has only a growth factor activity experimental annotation). The ROC-type analysis is illustrated in Figure 2.6. As with the results from the previous analysis, the impact of the truncation on all but level 1 appears minimal. Even at level 1 the results are comparable, and the quality of the results is superior to traditional pairwise approaches such as BLAST.

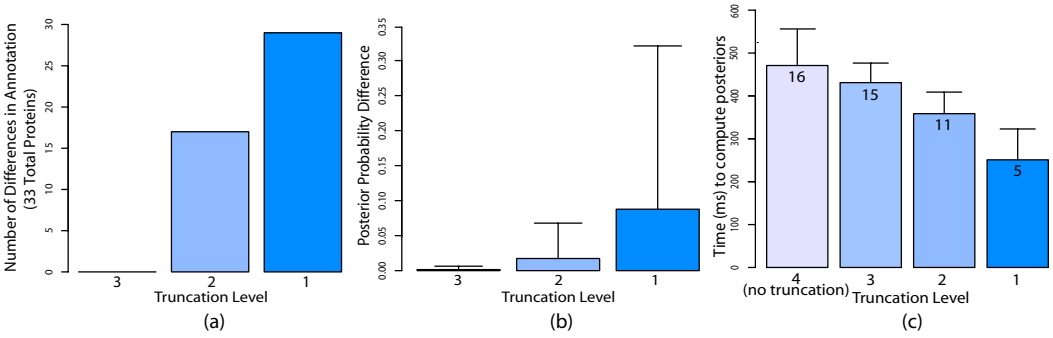

Figure 2.5: **Truncation approximation performance in the AMP/adenosine deaminase family.** Impact of truncation levels on the deaminase family. Recall there are four candidate functions for the AMP/adenosine deaminase family. Panel (a) shows the number of differences in molecular function predictions for every protein at the leaves of the phylogeny, truncating at each of the three possible levels for a maximum of 251 possible differences. This does not evaluate whether the predictions on the entire family of proteins were correct or not, only that the predictions matched. Panel (b) shows the mean L_1 (absolute value) difference between the approximate posterior probabilities and the exact posterior probabilities, including the standard deviation of that difference. This figure also is for proteins at the leaves of the phylogeny, and includes bars for each of the three possible levels of truncation as compared to exact computation. Panel (c) shows the average time to compute posterior probabilities for all levels of truncation (including no truncation), averaged over 10 runs. The numbers inside the bars in figure (c) indicate the number of rows and columns of the matrix Q.

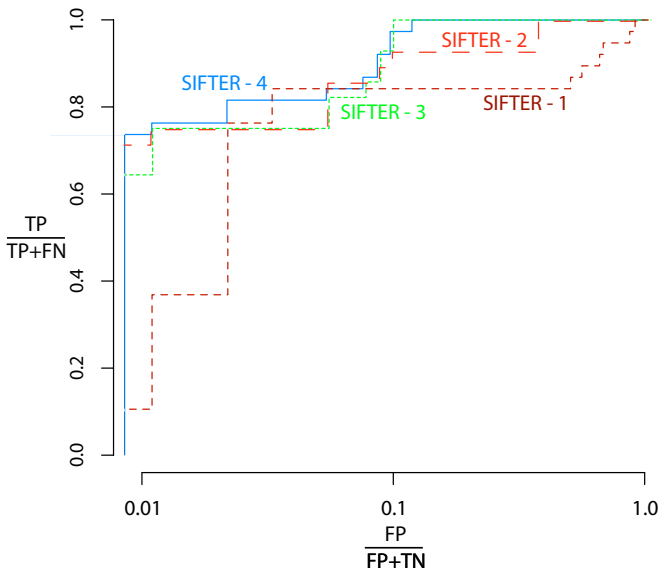

Figure 2.6: **Truncation approximation accuracy in the AMP/adenosine deaminase family.** Impact of truncation levels on cross-validation with expectation maximization. This figure shows the results of the ROC-type analysis on SIFTER leave-one-out cross-validation runs on the AMP/adenosine deaminase family of proteins. In the figure the curves are labeled SIFTER-TN where N is the level of truncation (4 is exact computation). Recall there are four candidate functions for the deaminase family. Levels 4, 3, and 2 all achieved the same accuracy (93.9%, or 31 out of 33), and level 1 achieved 90.9% accuracy (30 of 33) for leave-one-out cross-validation, where each run estimates the model parameter using EM. Note that the x-axis is on a log scale.

2.3.3 Sulfotransferase family

We applied SIFTER version 1.2 to the sulfotransferase family of enzymes (PF00685) from Pfam release 20.0. We used 539 proteins in this family, 48 of which have experimentally characterized molecular functions recorded in the GOA database. There are nine different candidate functions in SIFTER, eight of which are sulfotransferases acting on the specific compound involved in the enzymatic reaction, and the last of which is the non-specific label *nucleotide binding*. Generally, these enzymes are responsible for the transfer of sulfate groups to specific compounds. Researchers have shown their critical role in mediation of intercellular communication in multicellular organisms [Bowman and Bertozzi, 1999]. Human sulfotransferases are extensively studied because of their role in metabolizing steroids, hormones, and environmental toxins [Allai-Hassani *et al.*, 2007], and because they are biologically linked to aging [Feyzi *et al.*, 1998], Alzheimer's disease [Bongioanni *et al.*, 1996], and neuronal development and maintenance [Gibbs *et al.*, 2006], and in the past ten years have been studied as therapeutic targets. *Plasmodium falciparum*, the causative agent of malaria, makes use of its own sulfotransferase proteins as cell-surface receptors to enter into the host, thus these proteins are a target for malaria prevention drugs [Chai *et al.*, 2002]. Similarly, sulfotransferases in *Mycobacterium tuberculosis* are thought to be possible virulence factors [Mougous *et al.*, 2004], and thus are a potential target for tuberculosis drugs or vaccines.

The phylogeny for this family, showing the 48 proteins with experimental functional annotations from the GOA database, appears in Figure 2.7. This figure gives us some insight into molecular function evolution in the sulfotransferase family. Interestingly, all of the proteins in this family with experimental annotations from the GOA database are from vertebrate species. In general, these proteins are annotated as acting principally on a single class of compounds, thus, at a general level, they

appear to be single function proteins. Moreover, it appears in the phylogeny that each different molecular function most likely arose from a single mutation event, so there is no indicated parallel functional evolution in this family. Thus, this family tests SIFTER's ability to annotate a functionally diverse family, as opposed to its performance on a family with significant parallel functional evolution or multifunction proteins.

We see in Figure 2.7 that 18 of the 48 family members have only *sulfotransferase* annotations, which is not specific enough to be included in our studies. More precisely, all members of this family will be annotated with the general sulfotransferase function if we consider the SIFTER annotations implicitly propagated to this parent term via the "is a" edges in the ontology, so we focus instead on the sulfotransferase terms associated with specific compounds. Two of the remaining 30 proteins have a *nucleotide binding* molecular function annotation, and the remaining 28 have one of eight specific sulfotransferase functional annotations. Five of these specific sulfotransferase functions appear only once in the tree. Generally, there are three major functional clades in the pruned phylogeny in Figure 2.7: the aryl sulfotransferases at the top of the figure, the heparin-glucosamine 3-O-sulfotransferases in the lower half, and the N-acelylglocosamine 6-O-sulfotransferases at the bottom of the figure. The five singleton candidate functions within these clades necessarily have incorrect annotations in leave-one-out cross-validation in this particular setup of the experiments, as no similar experimental observation is available for transfer within the tree and the model parameters do not facilitate function mutations to unobserved molecular functions in the leave-one-out-type analyses.

We chose not to estimate parameters because the number of parameters for this family, based on its functional diversity, is larger than the number of available observations, making estimation unproductive. Instead, the parameter values were fixed to $\phi_{ii} = 0.5$, $\phi_{ij} = 1.0$ for $i \neq j$, $\sigma_{speciation} = 0.03$, $\sigma_{duplication} = 0.05$, and

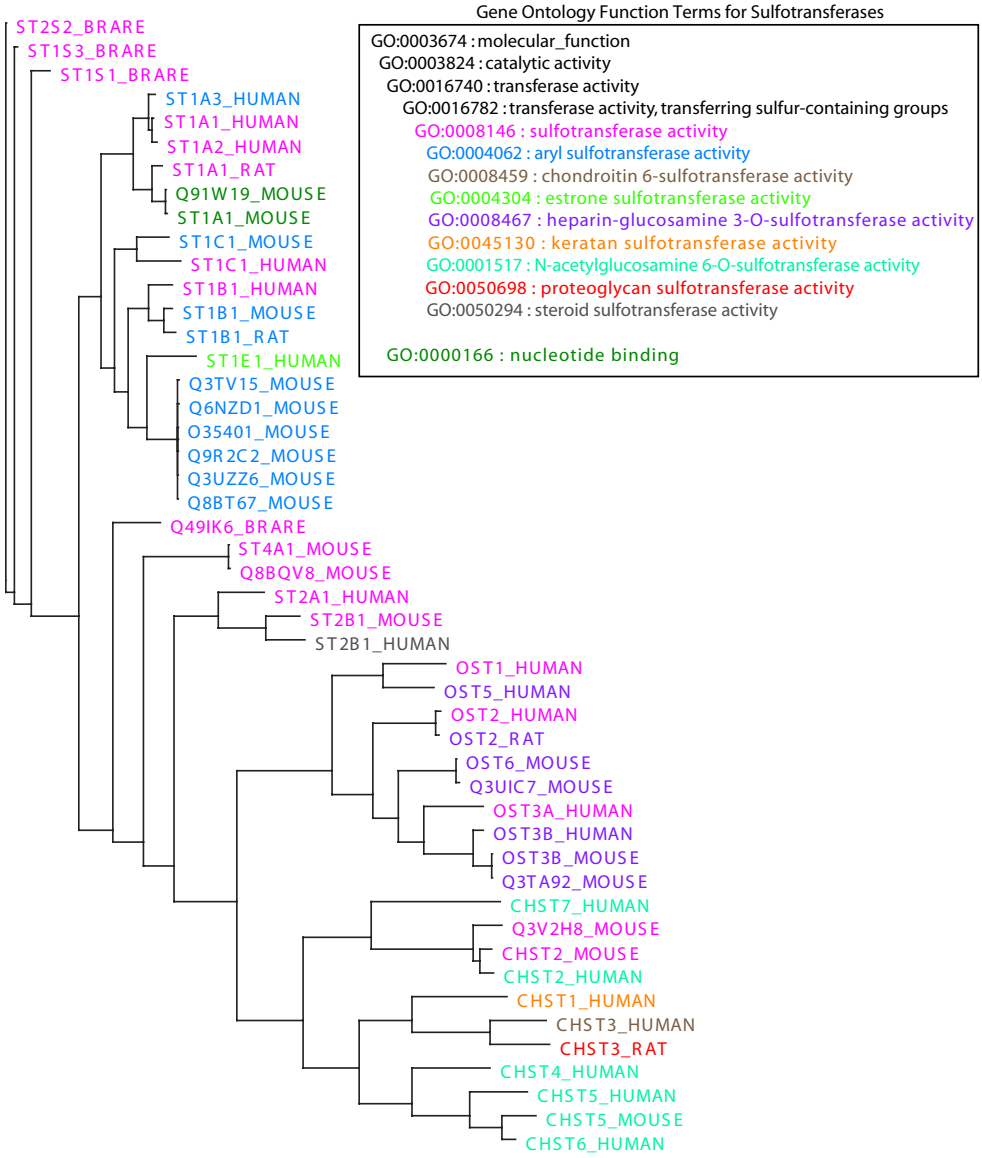

Figure 2.7: **Phylogeny of experimentally characterized sulfotransferase proteins.** The phylogeny of the experimentally characterized proteins found in the sulfotransferase family. We built this phylogeny by extracting the protein sequences for the experimentally characterized proteins from Swiss-Prot/TrEMBL, aligning them using the sulfotransferase HMM profile with `hmmalign` [Eddy, 1998], and reconstructing a phylogeny using PAUP* maximum parsimony [Swofford, 2001]. The 18 proteins with only the general *sulfotransferase* annotation (colored pink in the phylogeny) were not included in the results because their annotations were not specific.

$\alpha_i = 1.0$. Leave-one-out cross-validation, using exact computation of posterior probabilities (but not estimating parameters), yields 70.0% accuracy (21 of 30) when considering the 30 proteins with experimental annotations more specific than *sulfotransferases* and those with *nucleotide binding* experimental annotations. The SWISS-PROT identification numbers for the incorrectly annotated proteins from exact computation are ST1E1_HUMAN, ST2B1_HUMAN, CHST1_HUMAN, CHST3_HUMAN, CHST3_RAT, ST1A3_HUMAN, Q91W19_MOUSE, ST1A1_MOUSE, and Q8BT67_MOUSE, which include the five proteins with unique annotations (the first five on this list) as anticipated.

2.3.3.1 Power set truncation approximation results

We ran exact (non-truncated) leave-one-out cross-validation with a fixed set of parameters on the sulfotransferase family, and compared the performance with each of the eight possible approximations, one at each level of truncation. Figure 2.8 shows the non-truncated computation (labeled SIFTER-T9) versus three different levels of truncation: 6, 2 and 1. The ROC-type curves for the omitted levels of truncation were too similar to those of level 6 and non-truncated computation to be included. The figure makes clear that the impact of the truncation is minimal at all but level 1, and even then the impact is small.

Alternatively, we can compare the accuracy of the leave-one-out cross-validation, where non-truncated computation achieves 70.0% (21 of 30) accuracy. Levels 8 through 3 achieve the same level of accuracy as the non-truncated algorithm, missing the same set of proteins, whereas levels 2 and 1 achieve 66.7% accuracy (20 of 30). At level 1, SIFTER makes an incorrect prediction for the protein OST5_HUMAN, whereas level 2, SIFTER predicts Q8BT67_MOUSE correctly, but makes incorrect predictions for ST1B1_MOUSE and ST1C1_MOUSE.

Additionally, we measured the time associated with these computations. Fig-

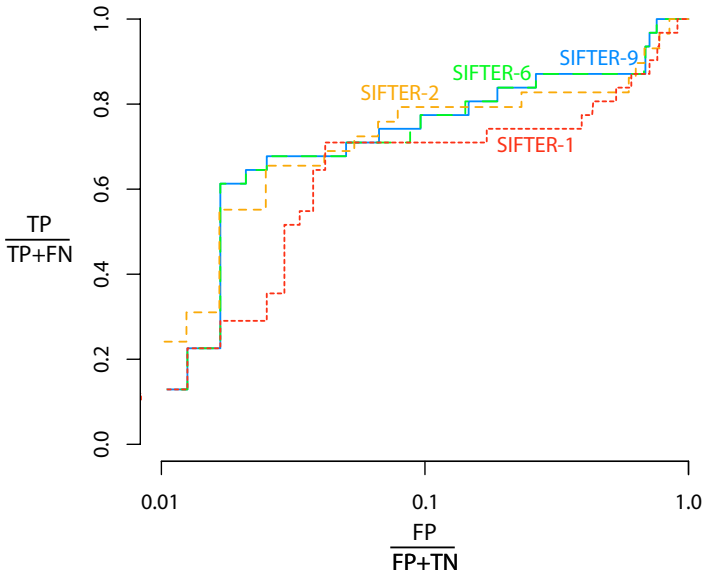

Figure 2.8: SIFTER **truncation approximation comparisons for the sulfotransferase family.** A comparison of different levels of truncation in the SIFTER truncation approximation for the sulfotransferase family of proteins. This ROC-type analysis shows the rate of false positives versus true positives as the acceptance cutoff varies from admitting no annotations to admitting all annotations. SIFTER's performance does not appear to degrade meaningfully from exact computation until truncation reaches level 1, and even then performance is still reasonable. Note that the x-axis is on a log scale.

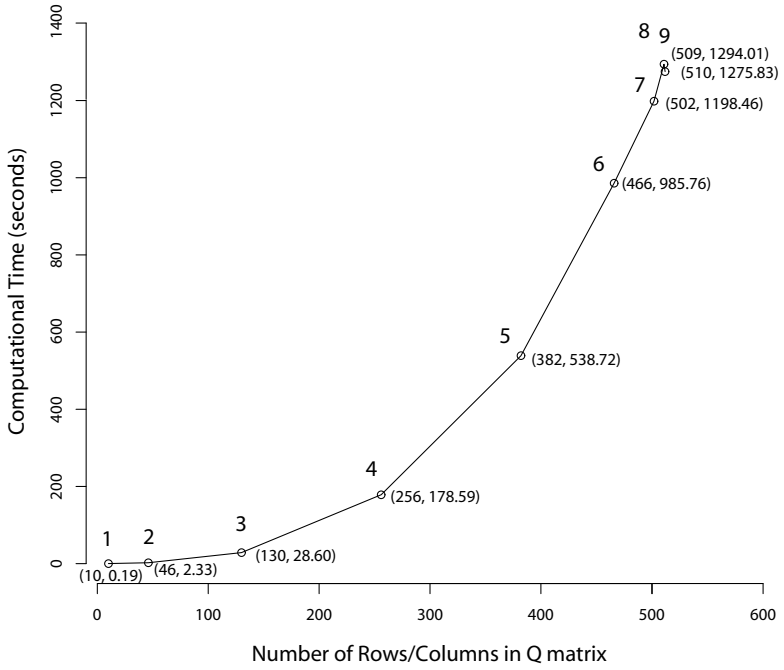

Figure 2.9: SIFTER **truncation approximation performance for the sulfotransferase family.** Computation time for SIFTER versus the size of the transition rate matrix. This graph illustrates how the time to compute posterior probabilities scales relative to the size of the transition rate matrix Q. There is a 50–500 times speedup in going from the complete matrix Q to a matrix truncated at $T = 3$ or $T = 2$ (where T indicates the level of truncation), with no meaningful loss in accuracy (see previous figure). The truncation level and (x, y) coordinates are included at each point for clarity.

ure 2.9 shows the tradeoff between the size of the transition rate matrix Q and the time of computation. Each of the time points is the mean of the time for each of the 48 cross-validation runs at each level of truncation (including exact computation). The truncation approximations improve the run time by a large margin – 500-fold in the case of of truncation level 2 – with minimal degradation in results.

2.3.3.2 Comparison with BLAST

BLAST achieves 50.0% accuracy (15 of 30) when run on the same subset of proteins in this family, run as described for the AMP/adenosine deaminase family. A ROC-type

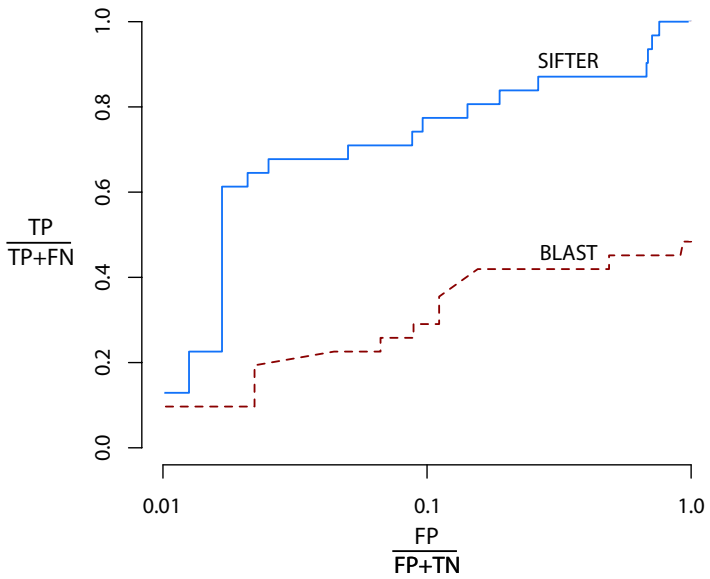

Figure 2.10: SIFTER-BLAST **sulfotransferase comparison.** A comparison of BLAST and SIFTER for the sulfotransferase family of proteins. This ROC-type analysis shows the rate of false positives versus true positives as the acceptance cutoff varies from admitting no annotations to admitting all annotations. SIFTER consistently dominates BLAST-type annotations under this criterion. Note that the x-axis is on a log scale.

analysis comparing the BLAST performance versus exact computation in SIFTER performance is shown in Figure 2.10, where SIFTER performs much better than BLAST at all levels of false positives. BLAST made correct predictions for six proteins that were missed by SIFTER, including ST1A3_HUMAN, ST1E1_HUMAN, Q8BT67_MOUSE, ST2B1_HUMAN, CHST1_HUMAN, and CHST3_HUMAN, four of which are proteins with unique function annotations. This opens the door to the possibility of using BLAST output when the BLAST-predicted function is not in SIFTER's set of candidate molecular functions.

2.3.3.3 Non-experimental annotations

As discussed above, five proteins could not possible be correctly predicted by SIFTER in the leave-one-out cross-validation, because they are the only protein with their par-

ticular experimental annotation. We investigated whether including non-experimental annotations might enable these to be predicted correctly in these experiments. Including non-experimental annotations as observations does not improve the results dramatically. We ran leave-one-out cross-validation on the set of proteins with experimental annotations and electronic (i.e., *IEA*, with a probability of correctness set to 0.2) annotations at truncation level 2, yielding 73.3% accuracy (22 of 30). This experiment predicted proteins ST1A3_HUMAN and ST2B1_HUMAN correctly, and CHST7_HUMAN incorrectly, as compared to the non-truncated experiments using only experimental evidence. Although ideally including electronic annotations would mitigate the problems associated with unique experimental annotations by including some of the same electronic annotations, for this diverse protein family we did not find this to be the case for all but one of the proteins with unique experimental annotations (ST2B1_HUMAN). This may be because, in certain families such as this one, GO experimental evidence is often for a more specific term in the GO hierarchy than the non-experimental evidence, thus there are still few or no examples of the appropriately specific term.

2.3.4 Nudix family

The family of enzymes termed Nudix hydrolases (PF00293) includes members found in all kingdoms: eukaryotes, bacteria, archaea, and viruses [McLennan, 2006; Koonin, 1993]. Characterized by the highly conserved 23 amino acid motif $GX_5EX_7REUXEEGU$ (where U is a hydrophobic residue and X is any amino acid), Nudix hydrolases are so named because of their initially discovered activity on **nu**cleoside **di**phosphates linked to some other moiety named **X**; a number of non-nucleoside substrates have since been identified [Koonin, 1993; Bessman *et al.*, 1996]. Activity has been reported on coenzymes (such as CoA, FAD, and NADH), nucleotide

sugars, dinucleoside and diphosphoinositol polyphosphates, capped RNA, and canonical and oxidized (deoxy)ribonucleoside di- and triphosphates. These functions suggest roles in nucleotide pool sanitation, the removal of toxic metabolic intermediates, mRNA stability, and signaling [McLennan, 2006; Galperin *et al.*, 2006]. While the Nudix motif forms a loop-α helix-loop structure that provides a general scaffold for coordinating cation binding and catalysis, residues that lie outside of the motif govern substrate specificity [Koonin, 1993; Mildvan *et al.*, 2005].

2.3.4.1 Family discussion

The Nudix family contains 3703 member proteins in Pfam release 20.0. We used the alignment for this family from Pfam (which aligned the Nudix motif), and reconstructed the phylogeny from the neighbor joining algorithm in PAUP* [Swofford, 2001] because of the large size of the family.

The Nudix family poses a particular problem because sequence diversity degrades the alignment quality of these homologous proteins. The low-quality alignment in turn degrades the quality of the tree reconstruction. A bootstrap analysis of the alignment used to build the maximum parsimony phylogeny in Figure 2.11 for the experimentally characterized proteins had an average bootstrap support of 38%, and all of the clades with more than 95% bootstrap support are noted by purple circles in the figure (there are only five such clades containing at most five proteins). This overall lack of support indicates how little information is contained in the alignment that is useful in reconstructing the sequence phylogeny. Attempts have been made to perform structural alignments and use this as a scaffold, but these too suffer from the family's diversity [Ranatunga *et al.*, 2004]. The impact on SIFTER is noise in the tree reconstruction with some consistent errors in annotation. Furthermore, the tree shows numerous examples of possible parallel functional evolution, and many of these proteins have multiple functions. Nonetheless, even with an inaccurate tree, SIFTER's

phylogenomic approach performs well in comparison to a pairwise method (BLAST) as shown below.

Our own manual literature search revealed 97 proteins with experimentally characterized function (some the same as the GOA experimental annotations), meaning that 2.6% of the proteins in this family have experimental annotations. The phylogeny of the subset of Nudix proteins with experimental evidence of molecular function is illustrated in Figure 2.11. There are 37 proteins in this family with experimentally-derived annotations in the GOA database, but comparing those annotations against our own literature search yielded a number of apparent mistakes in the GOA database, so we used only the annotations from our literature search. We also found the set of GO terms for this family incomplete, so we augmented the ontology with 98 additional molecular function terms. Furthermore, we found some errors and inconsistencies in the hierarchical structure, so we rearranged the hierarchy slightly for biochemical accuracy (reorganized or renamed terms denoted with an asterisk in the figure). Each of the added terms are found in the literature experimentally characterizing a member of the Nudix family, and are descendant from the *hydrolase activity* term. The complete reconstructed GO subgraph of functional terms is shown in Figure 2.12. We restricted our search and GO augmentation to experimentally characterized hydrolases. Details of this analysis will be provided elsewhere.

In Figure 2.11, the sequences were aligned using `hmmalign` [Eddy, 1998] with the Nudix HMM profile from Pfam release 20.0. The phylogeny was built using PAUP* maximum parsimony [Swofford, 2001]. This phylogeny is used for illustrative purposes only, and was not actually used in SIFTER (instead, the full neighbor joining phylogeny including all 3703 proteins was used in SIFTER). Function annotations were assigned to each protein from the literature by contrasting different levels of experimental evidence for a given hydrolase to prune the less specific substrates from a larger list of assayed compounds.

Figure 2.11: **Phylogeny of experimentally characterized Nudix proteins.** The phylogeny of the experimentally characterized hydrolase proteins found in the Nudix family. We built the phylogeny by extracting the protein sequences for experimentally characterized proteins from Swiss-Prot/TrEMBL, aligning them to the Nudix HMM profile using `hmmalign` [Eddy, 1998], and reconstructing a phylogeny using PAUP* maximum parsimony [Swofford, 2001]. The colored numbers in brackets after the protein name represent the functional terms associated with each protein, from Figure 2.12, determined from our own literature search. The purple dots represent clades that have greater than 95% bootstrap support, for a bootstrap analysis of a neighbor joining tree using the same alignment in Phylip [Felsenstein, 1989] with 100 replicates.

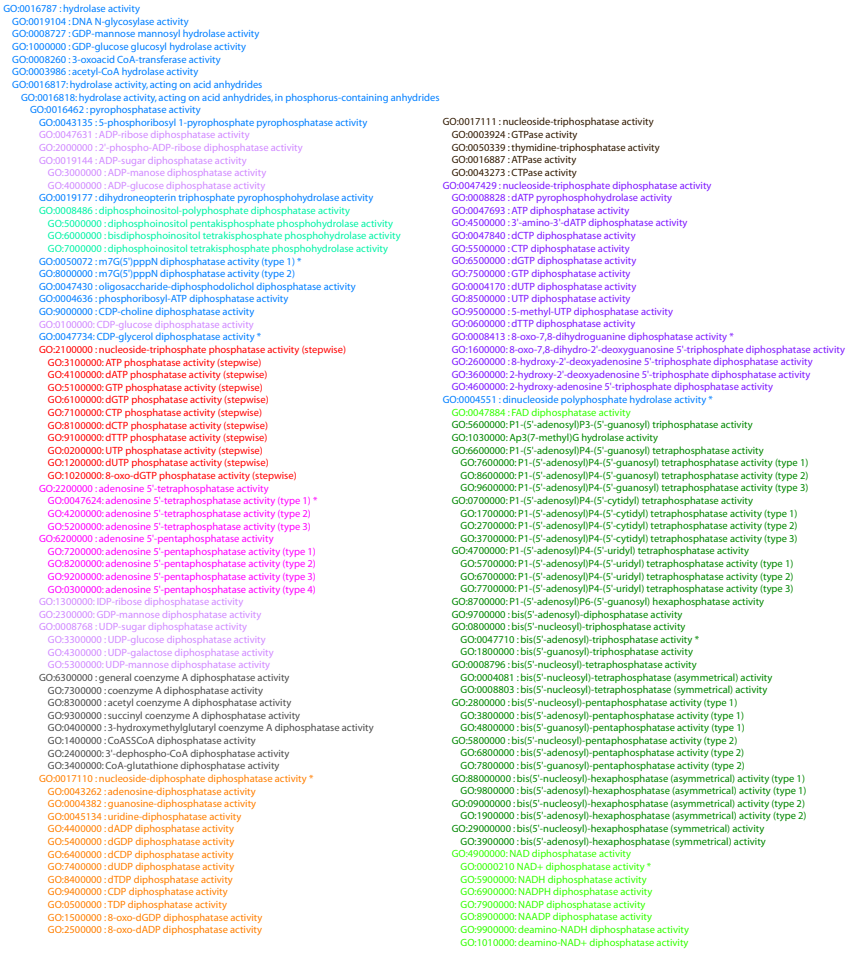

Figure 2.12: **Functional terms for the Nudix hydrolases.** Functional hydrolase terms that have been positively experimentally characterized in proteins from the Nudix family. This image captures the GO directed acyclic graph representing terms that are descendants of *hydrolase activity* in the molecular function ontology (noting that this particular subgraph is a strict tree, meaning that these nodes have only single parents). The terms with GO numbers greater than $1,000,000$ are ones that we have added to complete the spectrum of molecular functions that are experimentally found among the Nudix hydrolases. The terms with GO numbers less than $1,000,000$ are existing GO terms. Terms with an asterisk are existing GO terms for which we have either modified the name to reflect the chemical function more appropriately or reclassified within the hierarchy. The DAG can be read off from this image (here, a tree) by taking each term with more indentation than the previous term to be an ancestor of that term. The second column starts as an immediate ancestor of the pyrophosphatase activity term. The coloring is used as a tool to cluster the terms visually into groups with related substrates and corresponding related chemical activity.

The 66 candidate functions and large family size (compared to the other families studied here) produced a rich phylogeny with intriguing possibilities for further investigation. One observation is that many proteins with identical or similar functions cluster tightly in certain areas in the tree, in particular nucleotide-sugar diphosphatase (pink terms), diphosphoinositol polyphosphate diphosphatase (aqua terms), coenzyme A diphosphatase (gray terms), and diadenosine polyphosphate hydrolase activities (forest green terms). NAD diphosphatase activities are interestingly split into two clades, one of which is composed of proteins that are predominantly specific only for NAD-related compounds, while the other is made up of hydrolases that are also active on ADP-ribose and other dinucleoside polyphosphates. A grouping of mostly ADP-ribose diphosphatases in the middle of the tree is unique in that it clusters tightly, it is distant from other nucleotide sugar diphosphatases, and, moreover, within this clade the eukaryotic and bacterial/viral hydrolases are in two distinct groupings. In addition, most non-ADP-ribose diphosphatases cluster distantly from ADP-ribose diphosphatases.

A few particular proteins are worth noting. DIPP_ASFB7 is the only diphosphoinositol polyphosphate diphosphatase that does not cluster with other proteins of the same function, but instead is closely aligned with another viral hydrolase demonstrating quite different functions (Y06L_BPT4). Another protein of note is Q9RVP7_DEIRA, a nucleoside *di*phosphate diphosphatase that is closely related to three nucleoside *tri*phosphate diphosphatases, perhaps pointing to a similar catalytic mechanism for these four proteins.

There are 66 candidate functions for this family, so choosing a candidate function at random would yield on average about 1.5% accuracy for the single function proteins. Because there is a prohibitively large number of candidate functions for SIFTER with no truncation, we computed posterior probabilities approximately with truncation at one, yielding 67 possible functions (including none). Truncating at two,

as a comparison, yields 4356 possible functions or function pairs, and takes on the order of a week of time to compute posterior probabilities.

2.3.4.2 Results of cross-validation with SIFTER

We ran leave-one-out cross-validation on this family and computed the accuracy of SIFTER on the held-out protein with experimental evidence using a fixed set of parameters. Because of the size of the tree and the number of candidate terms, we did not attempt to estimate the large number of parameters for this family; instead, the parameters were set to the following: $\sigma_{duplication} = 0.05$, $\sigma_{speciation} = 0.03$, $\alpha = 1.0$ and $\phi_{ij} = 1.0$ for $i \neq j$ and $\phi_{ii} = 0.5$. Although these parameter settings are arbitrary and use little biological information other than crude intuition, they satisfy all of the parameter constraints mentioned in the Methods section. We set the truncation level to $T = 1$ for all of the Nudix experiments. SIFTER achieves 47.4% accuracy (46 out of 97 proteins annotated correctly) in this experiment.

With truncation at one, the 66 candidate functions have a matrix of size 67 by 67. The average time for computing posteriors on these machines was 146.78 seconds with a standard deviation of 0.62 second, as averaged over the 97 computations involved in the leave-one-out cross-validation.

2.3.4.3 Functional diversity in the Nudix family

The large functional diversity in this protein family is the main reason for difficulty in inferring molecular function. In this family, our data include five proteins (Q4U4W6, Q53738, Q81EE8, P32056, and O35013) with single, unique functions (i.e., they are the only protein in the tree to have that experimental annotation). In the case of protein O35013, it is labeled with four functions that are all unique to this family. From the perspective of the functional terms, as shown in Figure 2.13, most of them

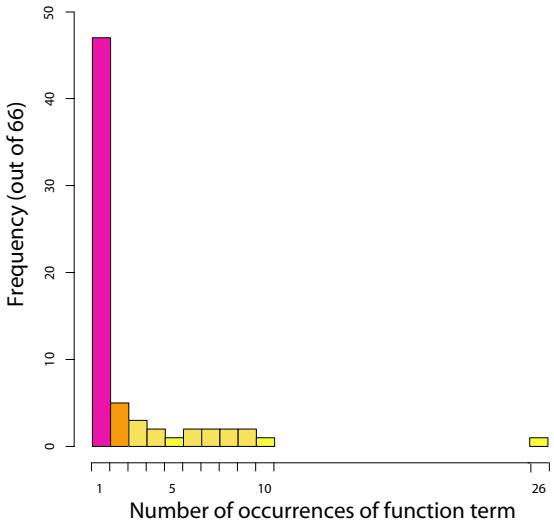

Figure 2.13: **Functional diversity in Nudix family.** Diversity of molecular functions in the Nudix family. This histogram illustrates the number of occurrences of each of the 66 different candidate functional terms in the 97 experimentally characterized proteins. Many of them occur only once; ADP-ribose diphosphatase occurs 26 times. This histogram represents the available characterizations and should not be used to interpret the relative counts of the functions in the entire family, as these counts may be skewed significantly by protein choice, assay difficulty, etc.

occur experimentally in this family once or twice, with the single extreme example of ADP-ribose diphosphatase activity occurring experimentally in 26 proteins in the family. The small number of proteins with common functional activity indicates that methods that predict molecular function via annotation transfer will encounter difficulty.

2.3.4.4 Generalizing functions

We wanted to examine the tradeoff between predicting molecular function at a more general level of the GO hierarchy and sensitivity. Within a family, we can selectively generalize some of the functional terms to improve our sensitivity when, for example, there exist characterization assays that provide a general screen for particular types of hydrolases. Although developing a method to automatically determine the

appropriate level of generalization is beyond the scope of this thesis, we manually generalized these functions to examine the impact on SIFTER's performance. Specifically, we generalized all of the terms in Figure 2.12 that are not colored blue, representing each set of non-blue terms by a single term. Viewed the blue terms as being already generalized, as they do not group according to biochemistry in a natural way and also occur closer to the root of the hierarchy relative to the set of terms we chose to generalize. After generalization there were 15 candidate molecular functions, 10 of which are generalized terms and the rest of which are original (blue) functional terms.

We ran leave-one-out cross-validation at truncation level 1, achieving 78.4% accuracy (76 of 97). Because the generalization reduced the diversity of this family extensively, we also ran leave-one-out cross-validation at truncation level 2, and it also achieved 78.4% accuracy (76 of 97). The ROC-type analysis for this experiment is shown in Figure 2.14, where SIFTER predicts 69.2% annotations correctly at 99% specificity. These experiments tell us that, when a biologist does not require a prediction to be as specific as the set of candidate functions, they may choose to trade term specificity for higher prediction accuracy.

2.3.4.5 Comparison with BLAST

We ran the Nudix sequences with experimental evidence through BLAST [Altschul *et al.*, 1990] on the non-redundant (nr) database with E-value cutoff 0.01. As in the two families discussed earlier, the BLAST output files were filtered through a keyword extraction script to convert the functional annotations from nr into GO terms (including our augmentations specific to the Nudix hydrolases), removing hits that were identical in species and sequence to the query sequence. This is the analog of a leave-one-out cross-validation experiment performed in SIFTER. BLAST uses all the proteins in the nr database with textual annotations describing any of the candidate functions, as opposed to SIFTER, which uses Pfam Nudix family member

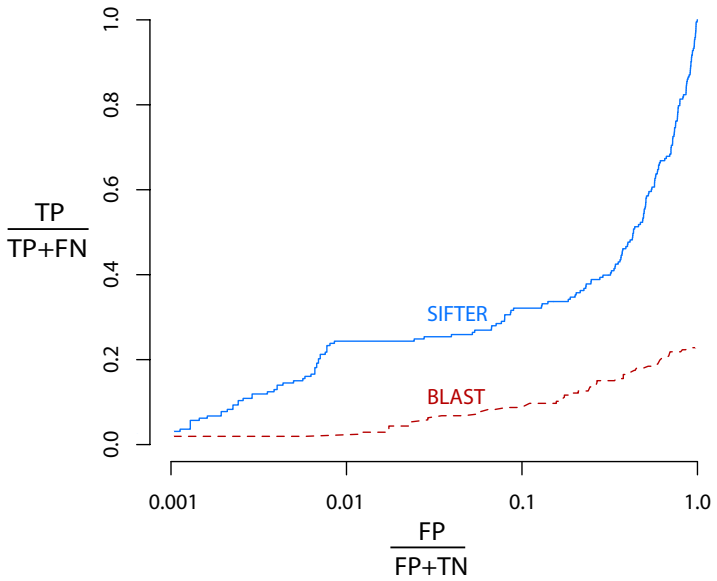

Figure 2.14: SIFTER-BLAST **Nudix comparison.** A comparison of BLAST and SIFTER for the Nudix family of proteins. The SIFTER-C line is the evaluation of SIFTER on the Nudix family using the 15 generalized Nudix hydrolase terms as experimental annotations. This ROC-type analysis shows the rate of false positives versus true positives as the acceptance cutoff varies from admitting no annotations to admitting all annotations. SIFTER consistently dominates BLAST-type annotations under this criterion. Note that the x-axis is on a log scale.

proteins with experimentally verified annotations found in the GOA database. We included all functional terms associated with the set of annotated proteins that share the most significant E-value, and consider a protein's BLAST annotation correct if one of the functions is in the set of experimental annotations for that particular protein. BLAST correctly annotated 33 of the 97 proteins at the level of the experimental characterizations, achieving 34.0% accuracy.

A more thorough comparison can be made between SIFTER and BLAST for the Nudix family using a ROC-type analysis, as in Figure 2.14. In the figure, which uses the results from SIFTER leave-one-out cross-validation and the BLAST output described above, we see that SIFTER outperforms BLAST at all levels of false positives. The SIFTER curve shows that, although functional diversity inhibits performance within this family, SIFTER still performs well, especially at low levels of false positives.

While overall accuracy appears fairly high for BLAST, the ROC-type analysis shows a general weakness of the BLAST results. Specifically, BLAST only identifies 23.3% of correct terms in the entire set of protein hits with E-value less than 0.01. In other words, including all predicted terms at any E-value (not just the most significant), only 23.3% of the Nudix terms show up at all in this set of BLAST predicted terms. This percentage is lower than the total number of correctly annotated proteins because 48 of the proteins have multiple functional terms that have been experimentally characterized. Although 34.0% of proteins had at least one of these terms correctly predicted, only 23.3% of the total set of terms was associated with the appropriate proteins, even at E-value cutoff of 0.01. At 99% specificity, approximately 2.4% of annotations are correct in BLAST, whereas for SIFTER, at 99% specificity, 24.4% of the annotations are correct.

2.3.4.6 Evaluating the value of observations

To evaluate SIFTER's sensitivity to data sampling, we left out multiple characterized proteins' annotations at each round of cross-validation. Specifically, we ran 2-, 3-, 5-, 10− and 20-fold cross-validation on this data set. In this type of cross-validation experiment, the data are randomly split into K disjoint sets (or *folds*), and the experiment is performed K times, leaving out one of the K subsets on each iteration during the posterior probability computation, and testing the accuracy of predictions on the held-out set. For 2-fold cross-validation, in which one half of the experimental annotations are removed at random for each run, SIFTER achieves 36.2% accuracy (35.1 of 97), as averaged over ten runs. For 20-fold cross-validation, in which approximately 5 of the experimental annotations are removed at random for each run, SIFTER achieves 46.1% accuracy (44.7 out of 97), as averaged over ten runs. As expected, as more evidence becomes available to SIFTER, the annotations improve up to a certain point, as shown in Figure 2.15. At 20-fold cross-validation, the accuracy is slightly less than leave-one-out cross-validation, quantifying the value of four additional observations out of the 97 total.

2.4 Basic SIFTER conclusions

We built the SIFTER methodology to work on large and functionally diverse protein families through a simple graphical model. We use a continuous time Markov chain and reduce the exponential time complexity through a straightforward but effective truncation of possible function combinations within each protein. SIFTER version 1.2 accuracy is comparable to SIFTER version 1.1, as shown on the AMP/adenosine deaminase protein family. The truncation approximation has excellent performance up to and including the final level of truncation on the AMP/adenosine deaminase

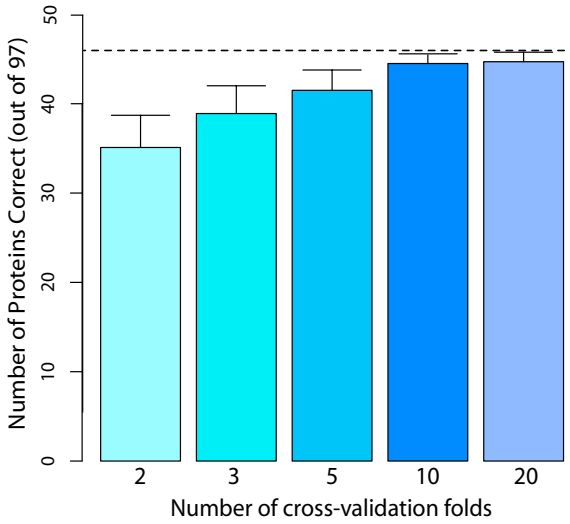

Figure 2.15: **Value of observations in SIFTER.** Number of correct annotations for SIFTER on the Nudix family of proteins across different numbers of folds. The x-axis of this figure represents five different partitions for cross-validation, from 2-fold to 20-fold cross-validation. The y-axis represents the average number of proteins for which SIFTER correctly predicted the function for each of the different cross-validation tests. The bars shown are the standard deviation for each partition. The dotted line at $y = 46$ represents the performance of leave-one-out cross-validation. All of the different partitions were run ten times.

family and the sulfotransferases. The sulfotransferases provide a good example of SIFTER's performance on a functionally diverse family of proteins that do not have any confounding evolutionary traits such as multifunction proteins or obvious parallel evolution. The new gold-standard data set for the Nudix family of proteins was built from a manual literature search. SIFTER performed well as compared to BLAST on the Nudix family, which is an exceptionally difficult family for function prediction methods.

Chapter 3

Choosing which Protein to Characterize

3.1 Introduction

The momentum behind automated methods for protein function annotation is primarily motivated by the costly, open-ended problem of experimentally characterizing protein molecular function. While DNA sequencing has become inexpensive and accurate, and finding protein structures is a well-defined (although costly) problem, experimentally characterizing protein molecular function requires some idea of the molecular function of a particular protein in order to select the appropriate assay, and even then the number of possible protein molecular functions is enormous. Proteins will sometimes perform multiple functions *in vivo*; examples include multidomain proteins such as heat shock protein 90 (HSP90) [Scheibel *et al.*, 1998], moonlighting proteins as in the aldolase proteins [Sriram *et al.*, 2005], or enzymes active on multiple substrates, such as many proteins in the Nudix family [Bessman *et al.*, 1996]. Thus identifying one type of molecular function does not rule out the possibility that

the protein has other functions with greater relative activity and that the identified function is not biologically relevant. It is possible that what we do not know about a protein family could be more important than what we do know; a recent paper suggested that single-gene disorders with Mendelian inheritance, including some metabolic disorders, may be due to mutations in moonlighting proteins for which we do not know the additional molecular functions of the responsible protein, and thus that protein would not be considered in the set of candidate genes [Sriram *et al.*, 2005]. We already use automated methods to perform molecular function annotation for biochemists on a large scale; how can automated methods help biochemists select which protein to characterize and which of the costly functional characterization experiments to perform (or not to perform)?

Experimental design (and the subfield of *active learning*) is a class of methods designed to answer this question. Designing a biological experiment requires defining all of the experimental variables, then selecting values for each of those variables. The experimental variables may include, for example, which proteins to characterize, the concentration of a particular substrate in a functional characterization experiment, or the duration of the measurements. Here we are concerned with selecting the minimal number of experiments that characterize a single protein with a specific molecular function activity such that the confidence of the functional predictions for the remaining unannotated proteins is maximized.

3.1.1 Active learning

Active learning is a subfield of experimental design in which an active learner interacts with the world to collect additional information in order to improve a supervised classification method. The framework of an active learning problem is as follows. An *active learner* is able to query the world and receive a response before outputting

its predictions. The active learner is responsible for selecting the query. The learner cannot impact the response to the query. There are two general methods for selecting the query: choosing to query the data with the *most uncertain* classification given the current model and parameters, and choosing to query the data that appear to be *most informative* with respect to a gain function.

Historically, many different algorithms and intuitions have been developed to decide how the query should be chosen. Query by Committee [Seung *et al.*, 1992; Freund *et al.*, 1997] puts a prior distribution over the classification hypotheses, and samples a set of classifiers from that distribution. Then the learner chooses to query one of the data points on which the classifiers maximally disagree. Uncertainty sampling [Lewis and Catlett, 1994] queries the data point that the current classifier is most uncertain about. This method is relatively easy to implement in many of the common classifiers, such as support vector machines where the point with the most uncertainty is the point closest to the margin.

Choosing an *optimal* query in the information-theoretic sense was formalized in the early 1990's [MacKay, 1992]. Three different query selection criteria were developed in this paper based on information theory for the purpose of maximizing the expected informativeness of the query responses. MacKay also noted a general drawback of active learning (and model-based methods more generally): the hypothesis space must be correct, or at least approximately correct, in order for these methods to effectively measure the value of the information. The first criterion that he presents selects the query where the classification has the highest variance given a set of observations. The second criterion selects a query in an area of interest that indirectly minimizes the mean marginal entropy of that particular class. The third criterion chooses queries such that the Gaussian distributions representing two different classes will tend to have maximally separated means and maximally different variances. The change in entropy of the parameters is used to validate each of these

approaches. None of these criteria directly address the goal of our active learner, which is that each query should, on average, maximally reduce the uncertainty about the functional predictions of the unobserved random variables.

Sequential design, which is a possible implementation of the active learning methods above, assumes that the model parameters are not known exactly and that multiple queries are possible [Pronzato, 2000]. The active learner makes a single query at a time, where each query is chosen based on the current data set and the current parameters, under the assumption that this may be the final query. In a single iteration, a data point is selected to query, the single query is performed, and the model parameters are updated based on the query response. This is repeated until no more queries are allowed.

We will use sequential experimental design methods to determine which molecular function assay to perform on which protein among a set of homologous proteins to maximize the value of that experimental outcome. The value of a particular assay is measured using mutual information, or the reduction in uncertainty for all of the remaining proteins' functions given the possible outcomes for the subsequent experiment. We attempt to find, at each iteration of active learning, the single experiment that maximizes the mutual information for a protein family, so as to be maximally confident in the SIFTER predictions for the unannotated proteins in the protein family using a minimal number of experimental assays.

3.2 Mutual information

The amount of information associated with a particular protein functional characterization given a set of observations can be described as the reduction in uncertainty about the remaining, unobserved random variables. Mutual information quantifies the dependence of two random variables X and Y, or how much knowing the value

of one variable reduces our uncertainty about the value of the other variable:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}.$$

From this equation, when X and Y are independent then $p(x,y) = p(x)p(y)$, and the value is zero. However, when they are the same event, and knowing X eliminates uncertainty about Y, then $p(x,y) = p(x) = p(y)$ and the mutual information function reaches its maximum value at the entropy of X (or, equivalently, Y).

Given the SIFTER model of how molecular function evolves in a protein phylogeny, we can select the assay that maximizes the mutual information within the tree. In particular, if we let X be the available experimental evidence, Y be a molecular function for a specific protein we might assay, and Φ be the model parameters, we must compute $p(X|\Phi)$, $p(Y|\Phi)$ (the prior probability of Y), and $p(X,Y|\Phi)$. If we note that $p(X,Y|\Phi) = p(Y|X,\Phi)p(X|\Phi)$ using the chain rule, the two equations on the right hand side are just the posterior probability of Y and the probability of X, which we can compute explicitly using our model. Thus, the computation of the mutual information is trivial: given the posterior and the prior probabilities of all of the leaf variables and the probability of the given set of protein functional observations, we can explicitly compute the mutual information for each possible experiment.

3.3 Evaluation techniques

Because of the prohibitively large cost of performing the actual molecular function characterizations, and the expertise involved in designing the assays and performing the experiments, we evaluated these methods using computational techniques. For each protein family used in validation, we initialized the data set to include one experimental observation (either *IDA*, *IMP*, or *TAS*) for each candidate function as-

sociated with that family. These observations were chosen at random from the set of possible experimental annotations. We computed posterior probabilities, conditioning on these M observations, and then ranked all of the possible experiments as observations based on an experimental design criterion. The highest ranking experiment with an actual experimental characterization was selected for inclusion in the set of observations, ensuring that a single characterization was not selected more than once (although a protein with more than one experimental observation would eventually have all of the experiments selected). At each iteration of the sequential experiment, we recorded the posterior probabilities of the proteins with unobserved experimental annotations. The experimental design criterion in our validation was mutual information criterion described above.

We also ran the same procedure where, at each iteration, we included an observation at random. In our experiments on a set of protein families, we compare the mutual information criterion against random inclusion, to determine which criterion reduces uncertainty maximally, while relying on the fewest experiments. We run each of these active learners 100 times for each family with random seed observations, to quantitatively compare the two active learning criteria.

3.4 Results on the AMP/adenosine deaminase family

Because of our previous work with this family and the relatively large set of experimental functional annotations available in our gold-standard data set [Engelhardt *et al.*, 2006], this family provides a baseline for the comparison of the mutual information criterion and random inclusion.

Figure 3.1 shows one feature for comparing the two types of active learners. In particular, at each cycle of the sequential active learner, as a single new observation is added to the set of observations for SIFTER, we computed the average posterior

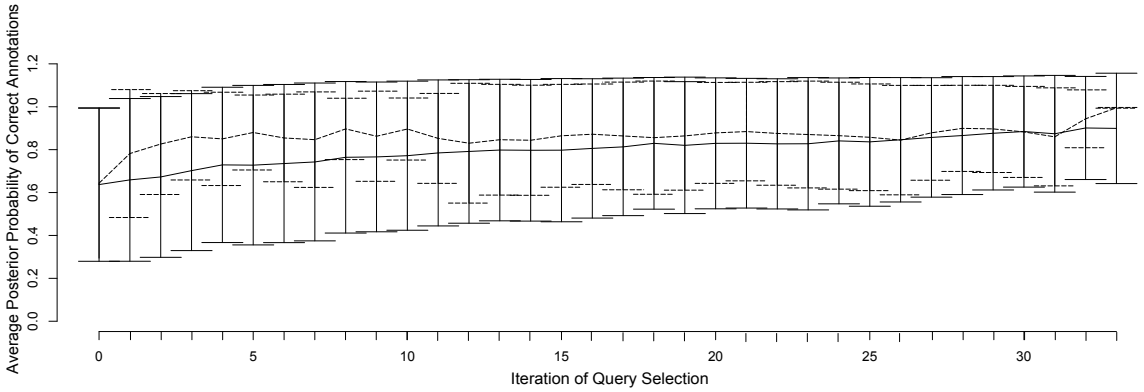

Figure 3.1: The average posterior probability of the correct functional annotations that have not been added as an observation in SIFTER's computation of the posterior probability on the AMP/adenosine deaminase gold-standard data set. This plot shows the average posterior probability of the correct, held-out functional annotations for the active learner selecting observations to add at random (solid lines) and for the active learner selecting observations to add using mutual information criterion (dotted lines), including the error bars for each representing the standard deviation.

probability of the held-out correct experimental functional annotations. The scenario depicted in the figure illustrates the power of the active learner using the mutual information criterion. It shows that, after including four additional functional characterizations, the average posterior probability of the held-out functional annotations is 0.860, whereas, including functional annotations at random, it takes 28 functional characterizations to reach an average posterior probability of 0.857 and 29 characterizations to reach an average of 0.866. In other words, with just four additional experiments selected using the mutual information criterion, we are equally certain about the remaining functional predictions in the tree as if we had randomly functionally characterized 28 additional proteins. This succinctly illustrates the power of active learning to select proteins to characterize.

3.5 Results on the sulfotransferase family

The sulfotransferase family is functionally diverse, and this particular family constrains the active learner significantly. Specifically, since there are nine candidate functions in all across this family, and five of those nine functions occur experimentally in only one protein in the family, those five proteins will always be among the set that are included *a priori* to the active learner as described in the experiment setup. Thus, the set of proteins for characterization is much smaller relative to the number of characterized proteins in the sulfotransferases than in the other families described here.

The results on this family are shown in Figure 3.2, which compares the posterior probabilities of the correct, held-out annotations after each iteration of the sequential active learner for the mutual information criterion versus random inclusion. Unlike the AMP/adenosine deaminase family, these results do not suggest that the active learner has chosen appropriate experiments for this family. The active learner using mutual information criterion is at most 0.05 greater than the average posterior probability using random inclusion, and the largest difference in the averages is 0.21 (at the second to last iteration), where using random inclusion generates the higher of the two average posterior probabilities. The standard deviations are comparable throughout the experiment. The active learner using mutual information criterion initially performs worse than random inclusion, and towards the final iterations again performs worse than random inclusion. This may be a result of insufficient positive experimental annotations relative to the functional diversity of this family. The five proteins with unique molecular function annotations reduce the average posterior probability overall, and selecting the queries to identify the remaining four functions correctly may require more (or different) observations than are currently available in this family. Thus this experiment may not reach the point where the mutual

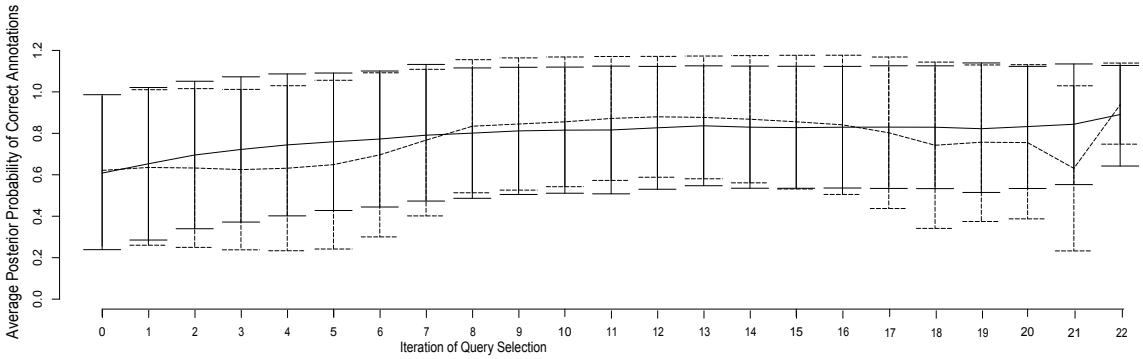

Figure 3.2: The average posterior probability of the correct functional annotations that have not been added as an observation in SIFTER's computation of the posterior probability on the sulfotransferase functional data set. This plot shows the average posterior probability of the correct, held-out functional annotations for the active learner selecting observations to add at random (solid lines) and for the active learner selecting observations to add using mutual information criterion (dotted lines), including the error bars for each representing the standard deviation.

information criterion shows much advantage over random inclusion.

3.6 Results on the aminotransferase family

The aminotransferase protein family (PF00155) is a member of the pyridoxal 5'-phosphate- (PLP-) dependent aminotransferase superfamily, responsible for converting an amino acid into its α-keto acid. This particular family is grouped according to specificity to two different amino acid substrates, aspartate aminotransferases (AATases) and tyrosine aminotransferases (TATases). Substrate specificity is actually defined as a preference for a particular substrate, meaning that the relative activity of a protein on one substrate is greater than the activity of that protein on another substrate, although *in vivo* it may actually act on both (or neither) of the substrates. In this particular family, the ratio of activities for the two substrates was, overall, much closer to one than many enzyme families (i.e., the actual substrate preference

in vivo was, in a few cases, not obvious) [Muratore *et al.*, 2007]. Furthermore, this family exhibits quite a few instances of observed parallel evolution, meaning that the preferred substrate mutates independently at more than one branch in the phylogeny for this family. We can observe this in the phylogeny of a subset of aminotransferases (Figure 3.3).

This family provides a different experiment than the previous two, because a subset of proteins in this family were selected to functionally characterize in order to maximally diversify the proteins that had functional information throughout the family [Muratore *et al.*, 2007]. This selection was performed in a different way, with a different goal, and from a different starting point than the SIFTER active learner, but it does provide an alternative comparison to the random approach. For completeness and consistency, we first present basic SIFTER results on this family, then we discuss the application of active learning to this family.

3.6.1 SIFTER results for the aminotransferase family

SIFTER was applied to this aminotransferase Iα family of proteins. We reconstructed a phylogenetic tree from the SATCHMO [Edgar and Sjolander, 2003] alignment of the family using PAUP* maximum parsimony [Swofford, 2001] and identified possible duplication events in the evolutionary history by reconciling the resulting tree using the Forester v.1.92 program [Zmasek and Eddy, 2001a] with a reference species tree from Pfam release 20.0 [Bateman *et al.*, 2002] for a subset of the aminotransferase superfamily PF00155 (specifically, the aminotransferase Iα family). We ran SIFTER including the set of functional characterizations in the gold-standard dataset created by Kathryn Muratore [Muratore *et al.*, 2007], as represented in the phylogeny in figure 3.3.

We ran SIFTER on the phylogenetic tree for the aminotransferase Iα family. Using

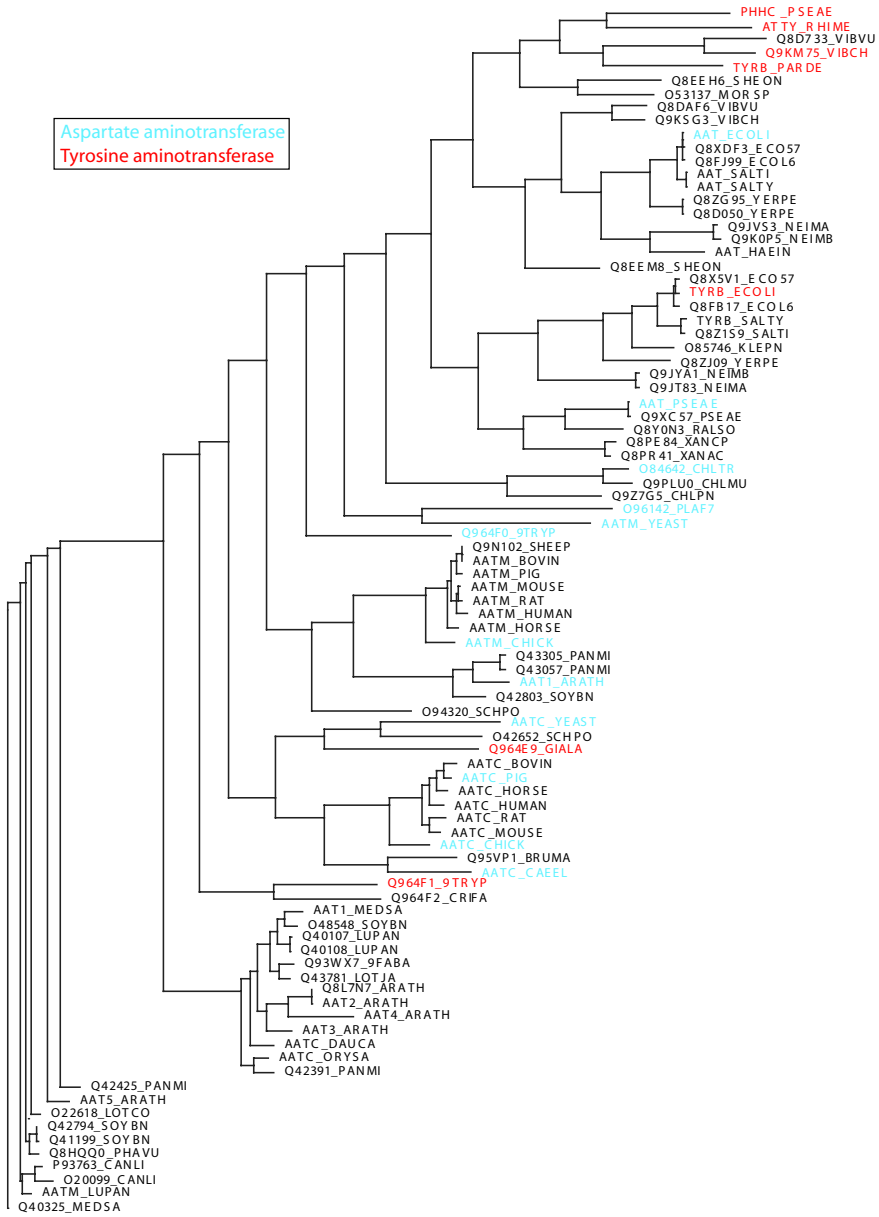

Figure 3.3: The phylogeny for the aminotransferase family, showing our gold-standard functional annotations for this family. The proteins with functional annotations found from a literature search and also experimentally characterized earlier [Muratore *et al.*, 2007] are colored based on a preference for one of two substrates, aspartate and tyrosine. Note that there appears to have been many more than one change of substrate preference in this tree.

the default parameter values for SIFTER, described in Chapter 2, SIFTER achieves 73.7% accuracy (14 of 19 correct predictions). The figure shows that there may be at least four (or possibly more) locations in the phylogeny where the substrate specificity changed independently from aspartate to tyrosine. However, the annotations are not dense enough, nor are the methods for gene tree/species tree reconciliation accurate enough, to attempt to co-localize duplication events with branches where substrate specificity may have changed. Although parallel evolution does not preclude the use of phylogenomic analysis, it does indicate that, for these sequences, molecular function does not consistently evolve in parallel with protein sequence; in other words, small changes in sequence do not always correspond to small changes in molecular function, and large changes in sequence do not always correspond to large changes in function. The specific amino acids that are responsible for determining substrate specificity in this family of proteins is discussed further elsewhere [Muratore *et al.*, 2007].

3.6.2 Mutual information criterion versus random inclusion

The comparison between using mutual information to select experiments and performing experiments at random is striking for this particular family. This is because a single protein, O84642_CHLTR, is always added last when the active learner with mutual information criterion was used, so the variance of the posterior probabilities for both the correct and the incorrect annotations are small (zero for the single remaining positive annotation from O84642_CHLTR) relative to the variance of the average posterior probabilities for random inclusion for both the correct and incorrect substrates. These comparisons are shown in Figures 3.4 and 3.5. Figure 3.4 shows that the average posterior probability of the set of correct, unobserved annotations when the experiments are added using mutual information is overall higher than adding experiments at random, despite the large amount of parallel evolution

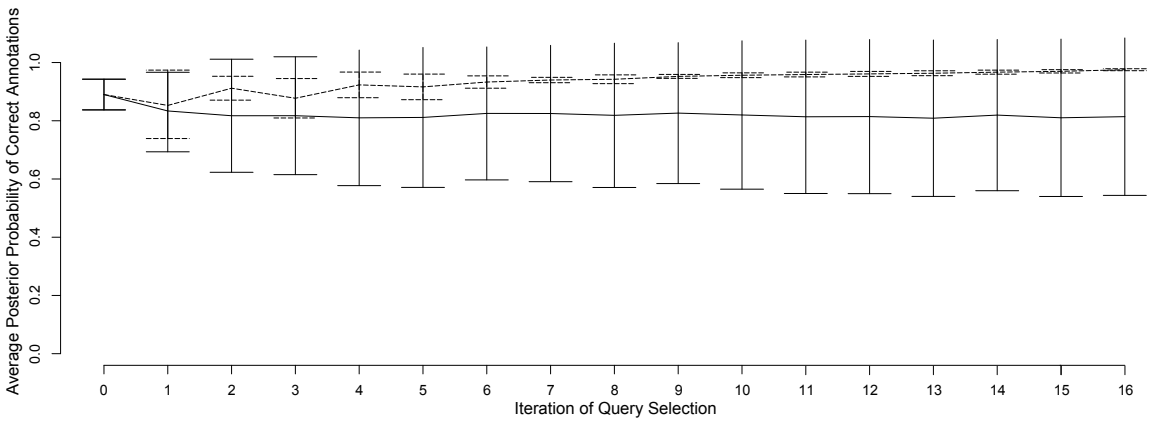

Figure 3.4: The average posterior probability of the correct functional annotations that have not been added as an observation in SIFTER's computation of the posterior probability. This plot shows the average posterior probability of the correct, held-out functional annotations for the active learner selecting observations to add at random (solid lines) and for the active learner selecting observations to add using mutual information criterion (dotted lines), including the error bars for each representing the standard deviation.

in this family. Figure 3.5 shows that, while the average posterior probabilities of the incorrect annotations for the random addition of experiments declines steadily, the same for the addition of experiments using mutual information criterion plateaus and then drops below the average for the random criterion after approximately half of the correct annotations are added. The variance of the posterior probabilities for the incorrect annotations is, except for the first position, smaller for the mutual information criterion than for random inclusion, reflecting an induced ordering of proteins using the mutual information criterion.

3.6.3 Prioritizing the proteins

To determine how each protein, on average, impacted the mutual information for the unobserved proteins, we found the average position (out of 17, since two proteins are added to the initial set of observations) that each protein was added to the set of

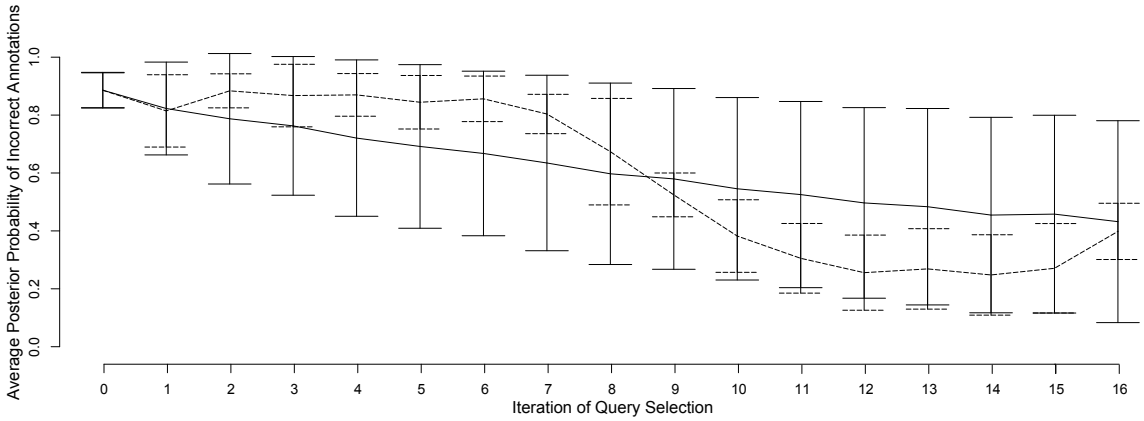

Figure 3.5: The average posterior probability of incorrect functional annotations. This plot shows the average posterior probability of the set of incorrect functional annotations for the active learner selecting observations to add at random (solid lines) and for the active learner selecting observations to add using mutual information criterion (dotted lines), including the error bars for each representing the standard deviation.

observations. In computing the average, we did not include the runs for which the protein was added to the initial observations.

The results for the aminotransferase family are dramatic as compared to the average position of inclusion, shown in Figure 3.6. Where the dotted line shows the average position of a protein in this rank order (using random inclusion, every protein's average position of inclusion was very close to the average, results not shown), there are four proteins above the average (including their standard deviations), indicating less informative proteins, and eight proteins below the line, indicating more informative proteins. In particular, the four proteins above the average are O84642_CHLTR, O96142_PLAF7, PHHC_PSEAE, and AATM_CHICK, where O84642_CHLTR is particularly interesting as it was added last in every case. The eight proteins that are below the average line are Q964E9_GIALA, TYRB_ECOLI, AATC_YEAST, AAT_ECOLI, Q9KM75_VIBCH, Q964F1_9TRYP, AAT1_ARATH, AAT_PSEAE.

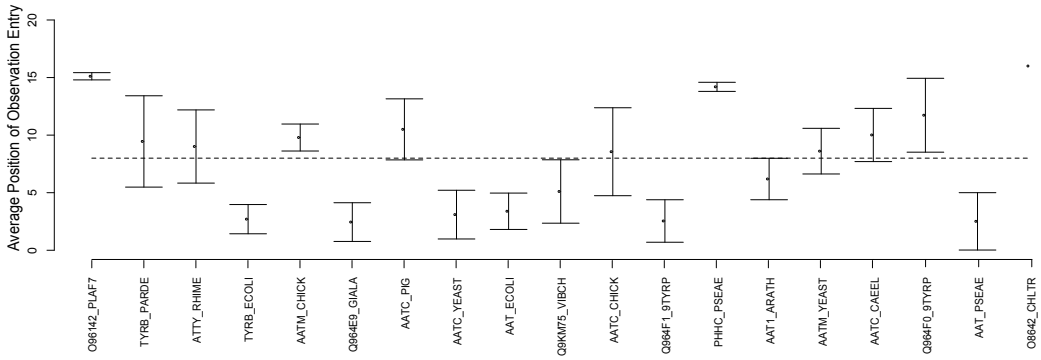

Figure 3.6: The average position of inclusion for each of the 19 aminotransferase proteins. The dotted line shows the average position of inclusion, and each protein has its average with standard deviation error bars for the 100 runs of the mutual information active learner. If a protein is included earlier, this is indicative of the protein being informative for the tree as a whole; if a protein is included later, this is indicative of the protein being less informative.

If we look at where each of these proteins occur in reference to each other in the aminotransferase phylogeny, we see that there are clusters of proteins whose particular observations are considered more or less informative. In particular, the two proteins with the least informative annotations are near each other in the phylogeny in terms of the number of branch traversals, namely O8642_CHLTR and O96142_PLAF7. These two proteins both have the same substrate preference (aspartate). On the other hand, two of the proteins with the most informative annotations, Q964E9_GIALA and AATC_YEAST, are also close together phylogenetically but have conflicting substrate preferences.

We can compare these "rankings" with rankings for a subset of these proteins through another method for finding a set of proteins to characterize [Muratore *et al.*, 2007], shown in Table 3.1. The Muratore scores compute the sequence similarity of this protein to other proteins in different subclades of the phylogeny. The higher the score, the greater the sequence distance, and thus the more informative the protein

Protein Name	Muratore Score	SIFTER Position
Q964E9_GIALA	17	2.4
Q964F1_9TYRP	15	2.5
Q9KM75_VIBCH	14	5.1
AAT1_ARATH	15	6.2
AATM_YEAST	21	8.6
AATC_CAEEL	10	10.0
Q96F0_9TRYP	21	11.7
O96142_PLAF7	25	15.1
O84642_CHLTR	31	16

Table 3.1: A comparison of the Muratore scores as compared to the average position of entry in the SIFTER active learning analysis. In this table, the proteins are ranked by SIFTER score. The scores have the reverse significance; in particular, for the Muratore scores, the higher the score, the more informative a protein should be relative to the available annotations, whereas for the SIFTER average position of entry, the lower the average position, the more informative the protein's particular annotation is.

is predicted to be. But the method assumes that both substrates are to be tested rather than ranking a single experiment, whereas the SIFTER active learner ranks the specific substrate experiment for the protein in question. To be clear, the mutual information criterion includes both the possibility of a positive and a negative experimental outcome, weighted by the probability of each given the current set of observed annotations. But we do not include negative experiments in the set of possible experiments for the active learner because currently SIFTER cannot incorporate the negative results of these experiments as observations in the posterior probability computation. If SIFTER could incorporate negative annotations as observations, we would have included the available negative characterizations in the set of possible queries for the active learner ([Muratore *et al.*, 2007] report most of these negative results).

Of note in the comparison is, first, that the protein O84642_CHLTR has both the highest score in the Muratore scores and the latest position of entry in the SIFTER

active learner, making the most informative and the least informative protein in this subset for the methods respectively. Using Spearman's rank correlation test, we get $\rho = -0.59$ which is not below the 0.05 level of significance (at -0.68) for the anti-correlation of these two measures. Thus, the two measures appear anti-correlated, but not at a significant level. The SIFTER active learner would also have to evaluate the expected value of the negative experiment for each protein to compare the two relative rankings more directly.

3.7 Experimental design discussion

Each of the families above shows a different quality of the active learner using mutual information criterion, and using these results, we may extrapolate on other, less well studied families how the active learner might perform, and more generally, the number of protein characterizations required for confident electronic prediction of functions for the remainder of the family.

The deaminase example is a straightforward example of the benefit of active learning. With only four additional, well-chosen experiments, we have equivalent confidence for functional predictions across the whole family as with 28 additional experiments at random. Although this family has proteins with multiple functions, this does not appear to be a limitation for the active learner.

Since SIFTER does not perform well in families with parallel evolution and minimal experimental evidence supporting the particular substrate preference in each functional clade, we do not expect active learning with SIFTER to perform particularly well on these families. As in the aminotransferase family, it appears that a fairly small set of characterizations will create equivalent confidence in the annotations for the remainder of the tree, although the confidence was not monotonically increasing. This is intuitive, as the locations of each of the mutations must be defined by at

least two different functional annotations, so characterizations of new functions in a particular region of the tree may add uncertainty to those regions temporarily. What this family does tell us, though, is that the choice of which proteins and functions to characterize is important, and some proteins are simply not important to characterize in high-confidence portions of the phylogeny (or equivalently because we are already so confident in the outcome of the proposed experiment).

For both the aminotransferase family and the sulfotransferase family, which has the largest functional diversity of the protein families used in validation, it is possible that this data set still lacks sufficient numbers of characterized proteins to find the benefit of the mutual information criterion over the random inclusion. In particular, perhaps it will take twice as many characterizations as are currently available to sufficiently define the locations of the functional mutations in these trees because of the parallel evolution. Since we have a limited set of annotations, we do not (and, in practice, will not) know when we have evidence for every instance of change of substrate preference in the tree, although we can attempt to estimate when are at a plateau in the uncertainty of the unobserved random variables.

Overall, using SIFTER's active learner with a mutual information criterion appears to work well in protein families with some functional diversity and possibly with multifunction proteins. Families with high functional diversity or parallel evolution may benefit less from a sequential active learning approach because of the large number of functional characterizations required to confidently predict molecular function for the unannotated proteins in the family. It is hard to quantify the benefit of active learning in such families using the small gold-standard data sets that are currently available.

3.8 Conclusions

In this work, we developed a sequential active learner using a mutual information criterion to select the next protein to functionally characterize (along with the specific function to characterize). We evaluated this method on three different protein families, the AMP/adenosine deaminase family, the sulfotransferase family, and the aminotransferase family, finding that in the AMP/adenosine deaminase family and the aminotransferase family, and less so in the sulfotransferase families, the active learner using mutual information criterion selects a small number of protein functional experiments required to reach the same level of functional prediction confidence relative to selecting these protein functional experiments at random.

Chapter 4

Fungal Genomes

4.1 Introduction

Fungal organisms may account for as much as 25% of the world's biomass [Gessner, 1997], and are important to study for a number of reasons. Fungi play a large role in our ecosystem, where they are responsible for decomposing and recycling vascular plants. Some species of fungi are either plant or animal pathogens. These include the rice blast fungus *Magnaporthe grisea* and human pathogens, which come from the genera Aspergillius, Candida, and Histoplasma. The human pathogens often infect immuno-compromised humans to cause serious diseases, but may also infect healthy individuals and are responsible for less serious diseases such as athlete's food or ringworm. Fungi are also useful for humans in a number of ways, namely in food and alcohol production through Baker's yeast, soy sauce production through *Aspergillius oryzae*, and the production of antibiotics. There are currently 46 fully sequenced fungal genomes, and more than 50 in progress. These represent a small fraction of the hypothesized number of existing fungal species, estimated to be somewhere between $712,000$ and 9.9 million [Hawksworth, 2004;

Schmit and Mueller, 2007].

Fungal species are interesting for scientists for a number of more technical reasons. Their genomes encode approximately 10,000 genes each, which makes them less than half of the size of the human genome. However, because they are eukaryotes, although they are structurally much simpler organisms (sometimes not multicellular), their cells look crudely similar to human cells in terms of the organelles and basic functionality. Because of this, the set of proteins critical to fungal organisms overlaps broadly with the set of proteins critical to animals. The species tree is fairly well understood, despite the sparse sampling of fungal species [James and et al, 2006; Fitzpatrick *et al.*, 2006]. One problem with reconstructing this tree using sparsely sampled species is that fungal species evolution is full of examples of convergent or parallel evolution, including at least four different times when the flagella were lost [James and et al, 2006], so morphological data does not guide the reconstruction.

4.2 Methods

The genomes from 46 different fungal species had been sequenced as of June 2006. Gene finding was performed in each of these genomes using a number of different methods, including GeneWise [Birney *et al.*, 2004], FgenesH+ [Salamov and Solovyev, 2000], and GLEAN [Mackey *et al.*, 2007]; for more specific details of their training and application, see [Stajich, 2006]. This resulted in 427,324 protein sequences from the different fungal genomes. We searched each of the resulting genes for Pfam domains using `hmmsearch` [Eddy, 1998] for the Pfam-A domains available in Pfam version 20.0 [Bateman *et al.*, 2002]. We aligned the set of fungal genomes corresponding to each Pfam domain by performing a hidden Markov model (HMM) alignment using `hmmalign` based on the Pfam-A HMM profile for that domain [Eddy, 1998]. We reconstructed phylogenies for each of the domains with PAUP* [Swofford, 2001], using

maximum parsimony with a BLOSUM50 matrix [Henikoff and Henikoff, 1992] when the size of the alignment file was less than 10000 kb, and neighbor joining when the size of the file exceeded that cutoff. We reconciled the trees against the species tree shown in Figure 4.1 using Forester version 1.92 [Zmasek and Eddy, 2001a]. Domains with fewer than four protein sequences from the 46 fungal genomes were eliminated. This resulted in 2800 phylogenies representing as many different sets of homologous protein domains within the fungal genomes.

Of the original set of proteins from the 46 fungal genomes, there were 236, 854 proteins that contained at least one Pfam-A domain and a family with greater than four members. We gathered molecular function annotations for each of the Pfam-A domains from the Gene Ontology Annotation (GOA) database in the following way. We ran BLAST version 2.2.4 [Altschul *et al.*, 1990] for each of the fungal proteins against the SWISS-PROT and TrEMBL fasta files, downloaded September 23, 2006 [Apweiler *et al.*, 2004], with an E-value cutoff of $1e-100$. We looked for exact hits in order to match the SWISS-PROT identifiers to the proteins in our set of fungal genomes. In particular, we selected a protein's exact SWISS-PROT or TrEMBL hit as the most significant hit from either database with E-value less than $1e-110$ and identical species. These identifiers enabled us to extract the molecular function annotations from the GOA database downloaded September 24, 2006 [Camon *et al.*, 2004].

4.2.1 Application of SIFTER

We applied SIFTER to each of these functionally annotated phylogenies. In particular, we computed posterior probabilities for all members of every tree, conditioned on the available evidence in the tree. We used both experimentally- and electronically-derived annotations as observations because of the sparsity of the experimental annotations. Trees with neither experimental nor electronic annotations were eliminated,

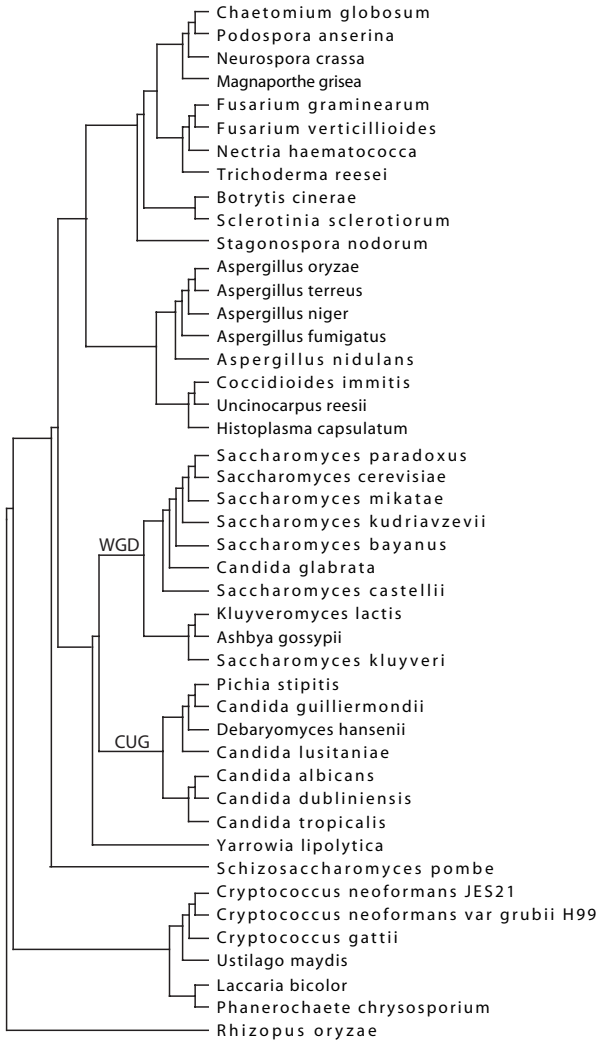

Figure 4.1: The set of fungal species with complete genomes sequences used in this study. The actual branch lengths were not estimated, and gene-species tree reconciliation does not use branch lengths. This tree was derived from tree reconstruction methods based on concatenating the sequences of 42 genes common to the set of fungal species, and then correcting for an instance of long branch attraction in the Aspergillus clade, as originally in [Fitzpatrick *et al.*, 2006]. We compared this tree to those found in two other sources [Stajich, 2006; James and et al, 2006] to build this consensus tree and to correctly insert the species in this study that were not in the original phylogeny.

as a SIFTER analysis of these families would be meaningless, leaving 2389 of the 2800 protein families for analysis, including 23 families with only a single candidate function (or 10 families if we only consider electronic annotations). We generated parameters for each of the families using the following defaults: the scale parameters $\sigma_{speciation} = 0.05$, $\sigma_{duplication} = 0.03$, the α parameters were all set to 1.0, and the Φ matrix had 0.5 on the diagonal, and 1.0 off-diagonal. The SIFTER runs were performed exactly for all proteins with fewer than nine functions; otherwise they were truncated at 2 for 9 to 19 candidate functions and at 1 for 20 or more candidate functions.

We ran SIFTER twice for each family. The first run included only experimental evidence for that particular family (i.e., including evidence with *IDA*, *IMP*, and *TAS* evidence codes), if available. The second run included all available evidence. Of the the 2389 families run on SIFTER, 552 had experimental evidence associated with one or more proteins, and 1837 had electronic evidence associated with one or more proteins. This yielded three possible types of experiments. The first type of experiment was on a protein family with both electronic and experimental evidence for which we predicted function given only the experimental evidence. The second type was on a protein family with both electronic and experimental evidence for which we predicted molecular function given both the electronic and experimental evidence. The third type was on a protein family with only electronic evidence for which we predicted molecular function given only the electronic evidence (i.e., there was no experimental evidence available). We report results for each type of experiment.

4.2.2 Limitations to methods

We see two limitations to the analysis performed here. For the less-well studied genomes, gene-finding is the primary source of the protein sequences as opposed to

any evidence of transcription or translation. The resulting hypothesized proteins will often differ from the actual expressed proteins in an insertion or deletion of multiple sequential amino acids. This is also true, to a lesser extent, for proteins from better-studied organisms hypothesized using espressed sequence tags, which identify transcribed, spliced nucleotide sequences. Thus, the BLAST searches for proteins with Gene Ontology (GO) annotations were not exact (i.e., requiring an E-value of 0.0), but instead used a low E-value ($1e-110$) cutoff. Since BLAST (as its name implies) is a local alignment tool, a sequence with an insertion or deletion of multiple, sequential base pairs relative to the query sequence will have a more significant E-value than a sequence with multiple, nonsequential mismatches. A second limitation is that the the families were built using Pfam-A release 20.0, where we used the profile HMMs for each of the Pfam-A domains to search for domains in each species' set of non-redundant hypothetical proteins. The implication of this is twofold: first, we will have missed identifying protein families that do not appear in Pfam-A or do not co-occur with a Pfam-A domain. Second, each family's proteins were identified by domain similarity, which is not always a guarantee of global homology.

4.3 Results

4.3.1 Fungal species

The collection of 46 fungal species used in this phylogenomic analysis is shown in Figure 4.1. The phylogeny of the species represents the best current understanding of the evolution of these species, despite such a sparse sampling of the species [Fitzpatrick *et al.*, 2006]. There are a few interesting clades in this tree that deserve a brief mention.

First, the fungal species in the clade labeled *CUG* are interesting in that they

use the codon CUG (i.e., CTG in DNA) to code for serine rather than leucine, as is the universal code in most other organisms [Laplazaa *et al.*, 2006; Ohama *et al.*, 1993]. This clade is interesting in that this particular characteristic appears to have evolved a single time in this tree, unlike many other fungal characteristics, such as sexual reproduction, multicellularity, and existence of a flagella. A second clade of note is the one labeled *WGD* in the figure, which includes species that are thought to have undergone a whole genome duplication (WGD) event [Kellis *et al.*, 2004; Fitzpatrick *et al.*, 2006] at the branch with the label.

Table 4.1 summarizes the set of fungal species used in this study, including the number of genes identified by gene finding methods and subsequently used in this study.

The average percentage of genes with a Pfam domain for the 46 fungal genomes is 57.0%, the highest percentage is *Schizosaccharomyces pombe* with 75.0% and the lowest percentage is 31.3% in *Laccaria bicolor*. After running these families through the SIFTER pipeline, which filters out families with fewer than four members and with no GOA database annotations, the average percentage of genes in a genome with results drops slightly.

4.3.2 A diverse set of protein families

We first investigated some of the properties of this landscape of protein families within the 46 fungal genome data set.

Of the 2800 phylogenies built, there were 2389 Pfam fungal families that had evidence from the GOA database and so were run on SIFTER. Of those families, the largest one was MFS_1 (Major Facilitator Superfamily) with 8729 member proteins, which is one of two types of transporter proteins that occur in all known organisms (the other being the ATP-Binding Cassette (ABC) superfamily, which has 1946 mem-

Species name	Genes	Genes in Pfam20	Predictions
Aspergillus fumigatus	9923	6077	8353
Ashbya gossypii	4726	3300	4306
Aspergillus nidulans	10701	6111	8820
Aspergillus niger	11200	7180	10604
Aspergillus oryzae	12075	7122	10464
Aspergillus terreus	11197	7066	10382
Botrytis cinerea	16448	5832	7676
Candida albicans	11128	7285	9795
Candida dubliniensis	6027	3960	5213
Candida glabrata	5272	3583	4731
Chaetomium globosum	11124	5642	7787
Candida guilliermondii	5920	3656	5024
Coccidioides immitis	10457	4778	6411
Candida lusitaniae	5941	3365	4513
Cryptococcus neoformans var grubii H99	7179	4371	5791
Cryptococcus neoformans var neoformans JEC21	6594	4334	5548
Cryptococcus gattii R265	6663	4017	5382
Cryptococcus gattii WM276	7196	4332	5382
Candida tropicalis	6258	3936	5076
Debaryomyces hansenii	6896	4230	5473
Fusarium graminearum	11640	6992	9535
Fusarium verticillioides	14179	7753	10896
Histoplasma capsulatum	9349	3908	4997
Kluyveromyces lactis	5331	3533	4620
Laccaria bicolor	20614	6454	8188
Magnaporthe grisea	12841	5848	8005
Neurospora crassa	9826	4938	6649
Nectria haematococca	16237	9731	13903
Podospora anserina	10443	4735	6368
Phanerochaete chrysosporium	12720	5462	7243
Pichia stipitis	5841	3551	4750
Rhizopus oryzae	17467	7874	10713
Saccharomyces bayanus	5490	2961	3929
Saccharomyces castellii	5706	2941	3858
Saccharomyces cerevisiae	5695	4127	5505
Saccharomyces kluyveri	5762	3305	4154
Saccharomyces kudriavzevii	6039	3553	4388
Saccharomyces mikatae	5702	3658	4747
Stagonospora nodorum	16597	6797	8928
Saccharomyces paradoxus	5512	3824	5069
Schizosaccharomyces pombe	5004	3751	4942
Sclerotinia sclerotiorum	14522	5711	7706
Trichoderma reesei	9997	5943	7926
Ustilago maydis	6522	4071	5548
Uncinocarpus reesii	8697	4632	6114
Yarrowia lipolytica	6666	4254	5659

Table 4.1: The 46 fungal species in the SIFTER study. Note that the number of predictions is slightly higher than the number of proteins in all of the fungal families, as the predictions count each prediction for each protein once (and for every family that a protein is a member of it receives a prediction), whereas the proteins in the families count proteins in multiple families only once.

bers in this data set). The average number of proteins in a fungal family was 165.3, with a standard deviation of 416.5, reflecting the enormous variability in size of these families. The total number of proteins in these families was 302,683, or 236,854 if we do not count proteins multiple times that appear in more than one family. There are slightly more internal nodes of the reconstructed phylogenies reconciled against the species tree shown above that are predicted to be associated with speciation events (52.6%), as compared to the number of internal nodes that are predicted to be associated with duplication events (48.4%). This is considerably more duplication events than one might expect, illustrating the large number of false positives associated with approximate tree reconciliation based on possibly erroneous trees.

On average, 35.3% of the proteins within a family had a matching sequence in SWISS-PROT or TrEMBL. Of those with database matches, on average 77.3% had annotations in the GOA database (meaning that 27.0% of the proteins in a family on average had at least one annotation from the GOA database). The somewhat astounding figure is the average number of annotations for an annotated protein from the GOA database was 3.5. This implies a "rich get richer" scheme within the GOA database (or perhaps biology in general) where a protein with a functional annotation appears more likely to accumulate additional (possibly redundant) annotations as compared to unannotated proteins. For experimental annotations from GOA, there are only 5336 experimental annotations in proteins in these families, meaning that the average number of experimental annotations for an annotated protein is 0.065. The average number of experimental annotations from the GOA database over the fungal families run on SIFTER is 1.7%. Of the 2389 families run on SIFTER, 23 have a single candidate function (when we include both experimental and electronic annotations), and 552 have experimental annotations for at least one of the member proteins.

4.3.3 SIFTER results

A few general observations about SIFTER's performance on these families will be mentioned. Then prediction results for the fungal species overall and individual families are discussed.

4.3.3.1 Timing

By far the most time-intensive step of the phylogenomic analysis was the phylogenetic tree reconstruction; this step would not have been as computationally intensive if we had chosen to use neighbor-joining trees at a much lower family size threshold. We computed the time for the SIFTER runs, given the reconciled phylogeny for a family and the associated GOA database annotations. When run on Dell Precision 390 Workstation computers with Intel Core2Duo 2.6 GHZ processors and 2 Gb RAM, the entire set of SIFTER runs took under seven hours. This could have been sped up considerably if we had truncated the computation of posterior probabilities in families with more than five candidate functions, as the majority of time was spent on the exact computations of families with five to eight candidate functions. As expected, the largest amount of time was spent on the families with greater functional diversity.

As a whole, the 1837 families with only electronic annotations from the GOA database took approximately 59.2 minutes, and the 552 families with experimental annotations from the GOA database took 47.8 minutes, whereas the same 552 families, including both electronic and experimental annotations, took 4 hours and 25.5 minutes. The average number of candidate functions is 9.4 in the families with both electronic and experimental annotations whereas the average number of candidate functions in the families with only electronic annotations is 3.7 and the average number of candidate functions in the families using only experimental annotations is 3.5, which explains why the average run time per experiment is over ten times higher from

the families using only electronic annotations to the families using both electronic and experimental annotations.

4.3.3.2 Identical versus consistent predictions

We can assess the accuracy of SIFTER using two different types of agreement between annotations and predictions, both represented using GO terms. We call a functional prediction *identical* or *correct* when it agrees exactly with the annotation term for that protein. We call a prediction *consistent* when it is a descendant of the GO directed acyclic graph (DAG) of the protein's functional annotation. In other words, a prediction is consistent with an annotation when the annotation is a more general descriptor of the predicted functional term, since the annotation does not rule out the possibility that the more specific term is a function of this protein. Because SIFTER gathers the set of candidate functions from the most specific annotations available, and predictions are from the set of candidate functions, we do not see SIFTER predictions that are ancestors of the functional annotations. Because of time limitations, we did not run leave-one-out cross-validation methods, but simply computed identical and consistent predictions relative to the annotations that were input to SIFTER. Therefore, the most appropriate measures of accuracy in these experiments are those that were trained using only experimental evidence and tested on the held-out electronic evidence.

While the average percentage of identical predictions for the runs including both experimental and electronic annotations was 56.3%, the average percentage of consistent predictions for the same set of runs was 90.9%. Similarly, the average percentage of identical predictions for the runs including only experimental annotations was 20.1%, whereas the average percentage of consistent predictions for those runs was 80.1%. The set of experiments for the families with only electronic annotations yielded 66.1% predictions identical and 96.4% predictions consistent with the electronic an-

notations. A more detailed account of these data appears in Figure 4.2, which shows, for each family, an overall histogram of identical and consistent predictions for all three types of experiments.

There are many more non-identical annotations than inconsistent annotations for all types of experiments. The reason for this is clear: the annotations in the GO DAG occur at many possible levels of the hierarchy. If we are only predicting functional terms at the most specific level of the hierarchy, then counting only exact term matches as *identical* means that predictions that are consistent with the current annotation (but not identical) are considered wrong. This seems like an inappropriate measure, and GO is designed specifically to enable more sophisticated measures of consistency. For the experiments including both experimental and electronic annotations, it is possible that as many as 34.6% of the annotated proteins in this data set on average were annotated at a less specific level than the SIFTER candidate functions, assuming that this reason alone accounts for the difference between the identical and consistent accuracy.

There are three basic sources of error, causing either non-identical or inconsistent predictions. The first is an error of omission, meaning that proteins with a particular molecular function have neither electronic nor experimental evidence in the GOA database. Second is an annotation error in the GOA database, where the functional annotation associated with a protein is incorrect. The third is an error in SIFTER's predictions. Each of the three experiments has some combination of these sources of error, and it would be useful to determine what percentage each source of error contributed to the non-identical or inconsistent predictions.

For all three types of experiments, the non-identical predictions signify places where the GOA annotation disagrees with the SIFTER prediction. This could be due to errors of omission. Biologically, this may happen for multifunction proteins, as in the AMP/adenosine deaminase family from Chapter 2, or moonlighting proteins.

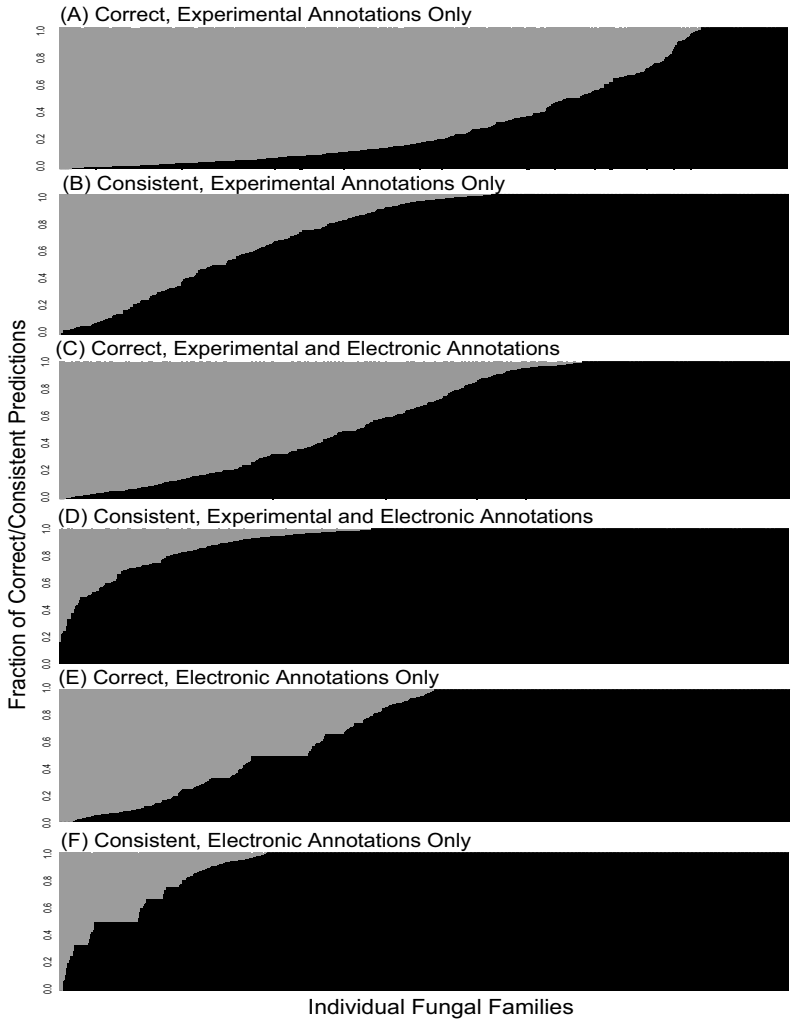

Figure 4.2: Six graphs showing the percentage of correct or consistent predictions for the three types of experiments. The percentage correct or consistent is colored black, whereas the percentage incorrect or inconsistent is colored grey. The bars do not reflect protein family size. Panel (A) shows the percentage of identical predictions for the experiments using only experimental annotations. Panel (B) shows the percentage of consistent predictions for the experiments using only experimental annotations. Panel (C) shows the percentage of identical predictions for the experiments using both experimental and electronic annotations. Panel (D) shows the percentage of consistent predictions for the experiments using both experimental and electronic annotations. Panel (E) shows the percentage of identical predictions for the experiments using the electronic annotations, for families without any experimental predictions. Panel (F) shows the percentage of consistent predictions for the experimental using electronic annotations, for families without any experimental predictions. Note in the final two that there is an interesting plateau around 0.5.

Another reason for non-identical predictions in this experiment is because of an error in SIFTER's predictions, perhaps because the evolutionary persistence of a function may be different than predicted by SIFTER. This is unfortunately a problem with the phylogenomic method more generally. Parallel evolution will often manifest as a short evolutionary persistence of a particular function, possibly at multiple locations in the tree. Incorrect alignments and incorrect phylogenies may also manifest similarly. Non-identical and inconsistent annotations may also be due to errors in the GOA database annotations. A manual phylogenomic analysis performed on those families may determine which source of error is responsible for the inaccuracies.

The experiments including only experimental annotations have higher rates of non-identity and inconsistency because the predictions based on experimental annotations were held-out from SIFTER's prediction. For these situations, it would also be desirable to determine which source of error the inconsistency or non-identity stems from.

4.3.3.3 Experimental annotations use more specific terms than electronic annotations

Using GO enables us to examine how specific or how general the terms used in annotations and subsequent predictions are. In particular, the root of the GO directed acyclic graph (at level 0, which is the term *molecular function*) is the most general term in the hierarchy, and as we traverse to subsequent levels in the hierarchy, the terms become more and more specific, for example, including substrate or cofactor information. The specificity of SIFTER's predictions in the fungal protein families reflects the specificity of different types of annotations in Pfam release 20.0. The histograms representing the average depth of the candidate functions across the set of fungal protein families is shown in Figure 4.3. The average level in the GO DAG of the candidate functions derived from experimental predictions is 6.52, whereas the

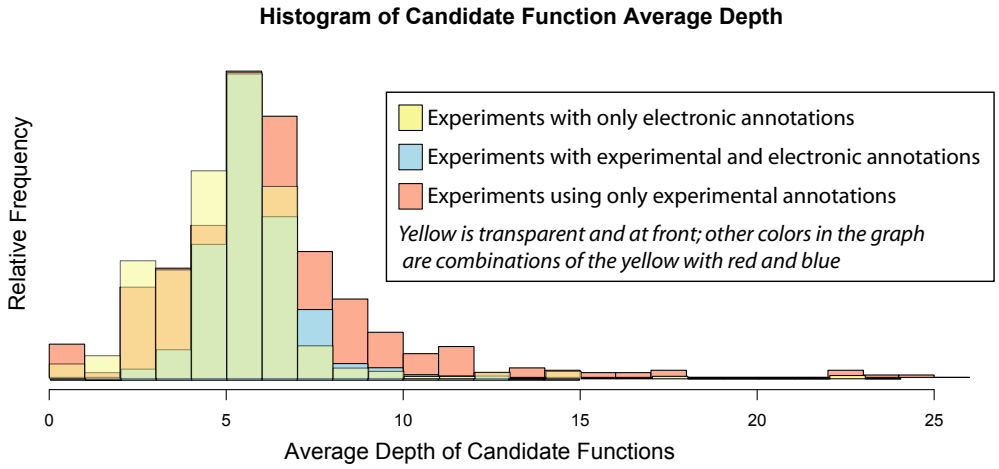

Figure 4.3: Histogram of average level of candidate functions for SIFTER runs with experimental annotations and experimental and electronic annotations. This figure shows the normalized histograms of the three different SIFTER experiment's candidate functions overlaid on top of each other. The histogram at the front (the yellow histogram, representing experiments with only electronic annotations) was made transparent so that the red and blue bars could be visible at the same time, appearing orange or green. The experiments using only experimental annotations dominate the histogram at average level six and larger.

average level in the GO DAG of the candidate functions derived from experimental and electronic predictions is 5.92, implying that experimentally validated functional annotations may be on average more specific than electronic annotations. In addition, there are 326 families in which the predictions from the set of experimental annotations are more specific than the predictions from all of the annotations, whereas there are 184 families in which the reverse is true. In the remaining 42 families, the two sets are at identical levels in the GO DAG.

If we consider the families for which the average level of the candidate functions from the experimental annotations indicates more specific terms than the average level of the candidate functions derived from both the experimental and electronic

annotations, this appears to occur because the experimental annotations are, in general, more specific than the electronic annotations. The glycotransferase-34 family includes this scenario; when only using experimental annotations, there are two leaves at level six of the GO DAG (alpha-1,6-mannosyltransferase activity and alpha-1,2-galactosyltransferase activity), and these annotation only occur in the proteins as experimental annotations. When the family is run with both experimental and electronic annotations, however, there are five candidate functions: the two functions from the experimental annotations, and three additional functions from the electronic annotations (*heme binding* at level four, *iron ion binding* at level six, and *monooxygenase activity* at level four). The reverse situation, where the average level of the candidate functions from the experimental annotations indicates more general predictions than is indicated from the average level of the candidate functions from both the experimental and electronic annotations, most often occurs as a result of omission, specifically when there were no experimental characterizations of a particular function within a family. There are many examples of this in this data set, as more than two-thirds of the families with annotations lacked any experimental annotations.

4.3.3.4 Overall functional diversity

We examine the diversity of functional predictions for each of the species in this fungal data set. For each protein in each family, we computed the fraction of SIFTER predictions descending from each of the molecular function terms at the most general level (the children of the node *molecular function*) of the GO DAG. We computed these figures for the experiments using both experimental and electronic annotations (Figure 4.4) and the experiments using electronic annotations with no available experimental annotations (Figure 4.5; the experiments using only experimental annotations give similar fractions to the experiments using both experimental and electronic, and are not shown).

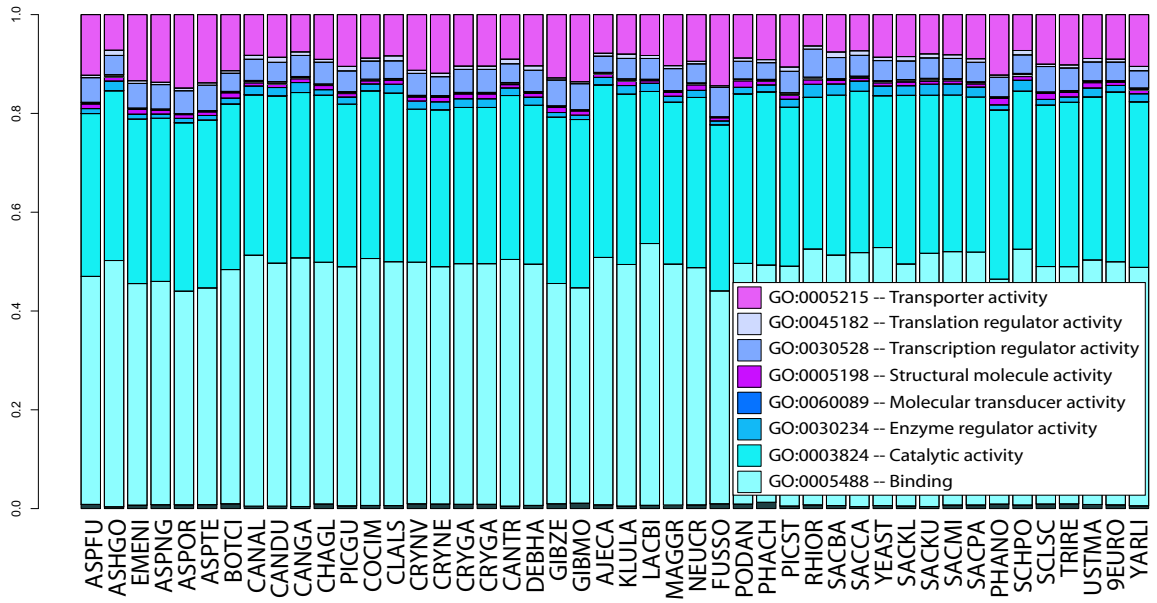

Figure 4.4: Overall functional prediction diversity for the fungal species. This figure shows the fraction of proteins found in each species that have SIFTER predictions in one of eight general molecular function categories. The names of each species are the first three letters of their genus concatenated with the first two letters of their species (e.g., *Aspergillus fumigatus* becomes ASPFU), with a few exceptions, including S. cerevisiae, which is abbreviated *YEAST*, and *Aspergillus nidulans*, which is abbreviated *EMENI*, both by SWISS-PROT convention.

The proportions from the SIFTER predictions using only electronic annotations show a slightly higher proportion of catalytic activity predictions relative to binding predictions, possibly reflecting the experimental bias that it is easier to show binding experimentally than many of the other functional classes.

The similarity between the fractions of each general molecular function for the proteins species does not trivially appear to correspond to the evolutionary similarity, with a few obvious exceptions. The fractions for *Candida albicans* and *Candida dubliniensis*, which have a recent common ancestor, appear very similar. In contrast, *Laccaria bicolor*, which appears to have a relatively large fraction of binding proteins, does not share this trait with *Phanerochaete chrysosporium*, the species most closely

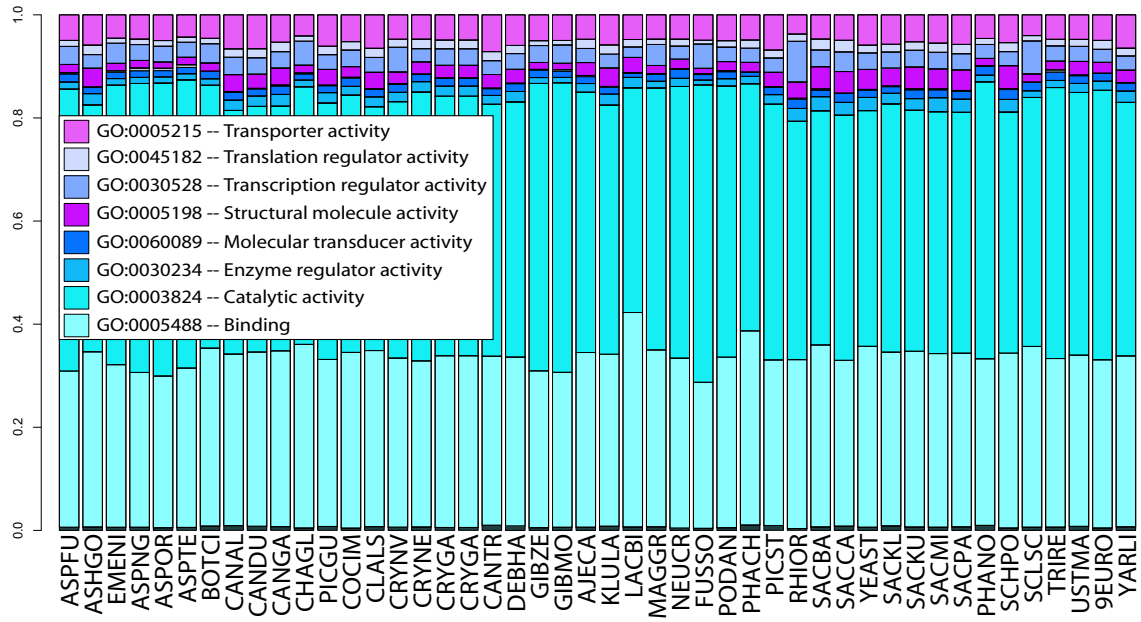

Figure 4.5: Overall functional prediction diversity for the fungal species for the set of protein families that only have electronic evidence. This figure shows the fraction of proteins found in each species that have SIFTER predictions in one of eight general molecular function categories. The names of each species are the first three letters of their genus concatenated with the first two letters of their species (e.g., *Aspergillus fumigatus* becomes ASPFU), with a few exceptions, including S. cerevisiae, which is abbreviated *YEAST*, and *Aspergillus nidulans*, which is abbreviated *EMENI*, both by SWISS-PROT convention.

related to it in this particular data set, although this could be an artifact due to the low number of genes found in *L. bicolor* or possibly a long branch length between them. Whether this lack of correspondence is an actual biological phenomenon, or whether it is a result of SIFTER errors or problems in gene finding, needs further investigation.

4.3.4 Specific families

We examined four specific families in this data set, two large and two small, to illustrate the power of function prediction using phylogenomic analysis. For each family, we describe the molecular functions within the family and the species that appear to have over- or under-represented numbers of proteins in these families to try to determine why duplications or deletions may have occurred in these organisms for this particular domain type. We also describe our literature reviews for these particular families and mention conclusions that may be drawn from our phylogenomic analyses of these families.

4.3.4.1 Tyrosine protein kinase family

The tyrosine protein kinase domain (Pkinase_Tyr, PF07714) has has 4772 members in this fungal data set and 24 candidate proteins (using only experimental data). There are two major clades in the protein kinase phylogeny, defined by their substrate specificity: one clade contains proteins that are serine/threonine specific, and the other contains proteins that are tyrosine-specific [Hanks *et al.*, 1988]. It is thought that tyrosine specificity is a more recent development, around the same time as multicellular organisms developed, and is important in cell-cell communication [Hanks *et al.*, 1988].

The two organisms that appear to have an over-representation of tyrosine protein

kinase domains are both pathogens. The average species has approximately 104 proteins with this domain in its genome. *Candida albicans* is a diploid fungus and a human parasite, causing thrush and, in immunocompromised patients, Candidiasis. *C. albicans* has 173 proteins with this domain in its genome. While normally a unicellular organism, *C. albicans* will change phenotype to become a multicellular organism (using both tyrosine protein kinase proteins [Cassola *et al.*, 2004] and SIR2 proteins [Perez-Martin *et al.*, 1999], also over-represented in the *C. albicans* lineage) in response to environmental factors [Cassola *et al.*, 2004]. A recent study showed that doubling the concentration of tyrosine protein kinase proteins reduces the growth rate of cells by two-fold [Cassola *et al.*, 2004], which does not explain why this apparently toxic protein recently appears to have had a large number of duplications in this organism.

No additional information was found on the role of tyrosine protein kinase proteins in the multicellular plant (and opportunistic human) pathogen *Rhizopus oryzae*, or any indications for why there appears to be an overrepresentation (244 copies) of the tyrosine protein kinase domain in this organism.

In the SIFTER analysis of this family, it appears that the fraction of proteins with predictions for each of the candidate functions was in general similar to the fractions of proteins specific to both *R. oryzae* and *C. albicans* with predictions for each of the candidate functions. A possible exception is that the percentage of proteins with protein histidine kinase activity predictions was slightly lower in both of these species than in the family as a whole. *R. oryzae* had 5.3% and *C. albicans* had 4.6% of their proteins predicted to have protein histidine kinase activity, as compared to the overall family that had 10.5% of proteins predicted to have protein histidine kinase activity. Overall, we can conclude that the recent duplications in both of these families do not appear to have occurred in a specific functional clade.

4.3.4.2 Major facilitator superfamily

Major Facilitator Superfamily (MFS) domain (MFS_1, PF) has 8729 members in this fungal data set and 31 candidate molecular functions using only experimental evidence in SIFTER. MFS proteins and the ABC transporters are the two types of transporters that occur in almost all organisms.

MFS proteins have the function of transporting potentially toxic substances out of the cell, but also for pathogenic species such as some of the fungal species in this set, these proteins can also facilitate the transport of toxins into the host cells.

The average number of proteins with this domain in a fungal genome in this data set is 190. It appears that there is significant over-representation of this domain in two clades of the fungal tree. Specifically, MFS appears to be over-represented in the three species *Aspergillus niger* (471), *Aspergillus terreus* (442), and *Aspergillus oryzae* (517), all forming a single clade. MFS is also over-represented in a second clade with three species: *Nectria haematococca* (674), *Fusarium verticillioides* (482), and *Fusarium graminearum* (379). Furthermore, this domain appears to be slightly under-represented in two species, *Laccaria bicolor* (121) and *Rhizopus oryzae* (125), but it is possible that this could be false negatives during the gene finding step.

The large amount of redundancy of MFS and related ABC and multi-drug resistance (MDR) protein domains has been documented in the Aspergillus genera, and is thought to be a cause of its development of resistance to a drug used for these human pathogens, triazoles, as these proteins have been found to reduce the overall intracellular accumulation of the drug [Ferreira *et al.*, 2005].

In the SIFTER predictions for the MFS proteins in two of the three of these *Aspergillus* species, it appears that there is a slight overabundance of predictions for spermine transmembrane transporter proteins (13.9% in *A. oryzae* and 16.6% in *A. niger* versus 10.0% in the family as a whole). In fungal organisms, spermine trans-

membrane transporters generally inhibit cell growth by excreting polyamines, which are charged molecules that are necessary for different reactions in the cell, but are cytotoxic in high concentrations [Tachihara *et al.*, 2005]. A number of studies have found that proteins with this function also play a role in salt ion transportation [Porat *et al.*, 2005]. In general, though, it appears that the recent duplications were found in many places in the phylogeny, and not in one specific clade.

The garden pea pathogen *N. haematococca* and the corn pathogen *F. verticillioides* are both well-studied pathogens, for which a great deal of information about how they infect their host is known. Although MFS proteins have been studied specifically in *F. verticillioides* [Lopez-Errasquin *et al.*, 2006], it is still not clear why this particular set of pathogens has duplicated the MFS domain so frequently. One possible thought is that a specific feature of these plant pathogens is that they are only pathogenic on a very limited set of specific host plants [Straney *et al.*, 2002], thus it is possible that they have duplicated the proteins required to transport the pathogens to the host in order to evolutionarily discover the optimal way to infect the host.

In the SIFTER predictions for the MFS proteins in these two species, it appears that there is no difference in the proportion of function predictions relative to the proportions for the MFS family as a whole. The largest difference is in spermine transmembrane transporters, which is predicted for 6.8% of the *N. haematococca* proteins and 7.2% for the *F. verticillioides* proteins, as compared to 10.0% for the full set of fungal species. In other words, it appears that the recent duplications were found across the phylogeny, and not in clades with specific molecular functions.

4.3.4.3 Spermine/spermidine synthase

Phylogenomic analysis of the spermine/spermidine synthase family explores some questions about whether these particular functions are essential to the fungal organisms. In this anaylsis, we are assuming that both the reconciled phylogeny and

the SIFTER predictions are correct. Spermine/spermidine synthase proteins are used in the cell to synthesize spermine and spermidine. In particular, polyamines such as spermidine are required for cell growth, and spermidine synthases are the proteins responsible for spermidine synthesis. There are two experimental annotations in this tree, both for proteins found in *S. cerevisiae*, and one of which is spermine synthase activity (GO:0016768), the other of which is spermidine synthase activity (GO:0004677). The phylogeny is shown in Figure 4.6, with the portions of the tree that SIFTER predicts as spermine synthase and spermidine synthase colored appropriately.

The phylogeny for this family shows that the two recent duplications of this particular protein domain in *C. albicans* may have occurred in the two different functional clades of this family phylogeny. It also appears that the duplication events may have happened very recently, as the sequence similarity of both sets of inparalogs is high. Because we performed an analysis on the set of nonredundant proteins for each of these species, though, we can be confident that these are not, in fact, the same protein in different strains of *C. albicans*.

There are a number of papers on whether or not these particular proteins are essential for the organisms. It appears that spermine synthase is not essential in *S. cerevisiae* [Hamasaki-Katagiri *et al.*, 1998], but spermidine synthase is essential for growth in both *S. cerevisiae* [Hamasaki-Katagiri *et al.*, 1997] and *S. pombe* [Chattopadhyay *et al.*, 2002]. The latter of these papers also indicates that the SIFTER prediction for SPEE_SCHPO is correctly spermidine synthase. This is interesting in terms of the phylogeny; since it appears that there are the same number of inparalogs (i.e., recent duplication events) in this tree in the essential spermidine clade (one inparalog, the *C. albicans* proteins) as compared to the spermine clade (one inparalog, the *C. albicans* proteins), it may be possible that there is no dosage requirement for the particular protein, as it appears that redundancy of the essential protein is al-

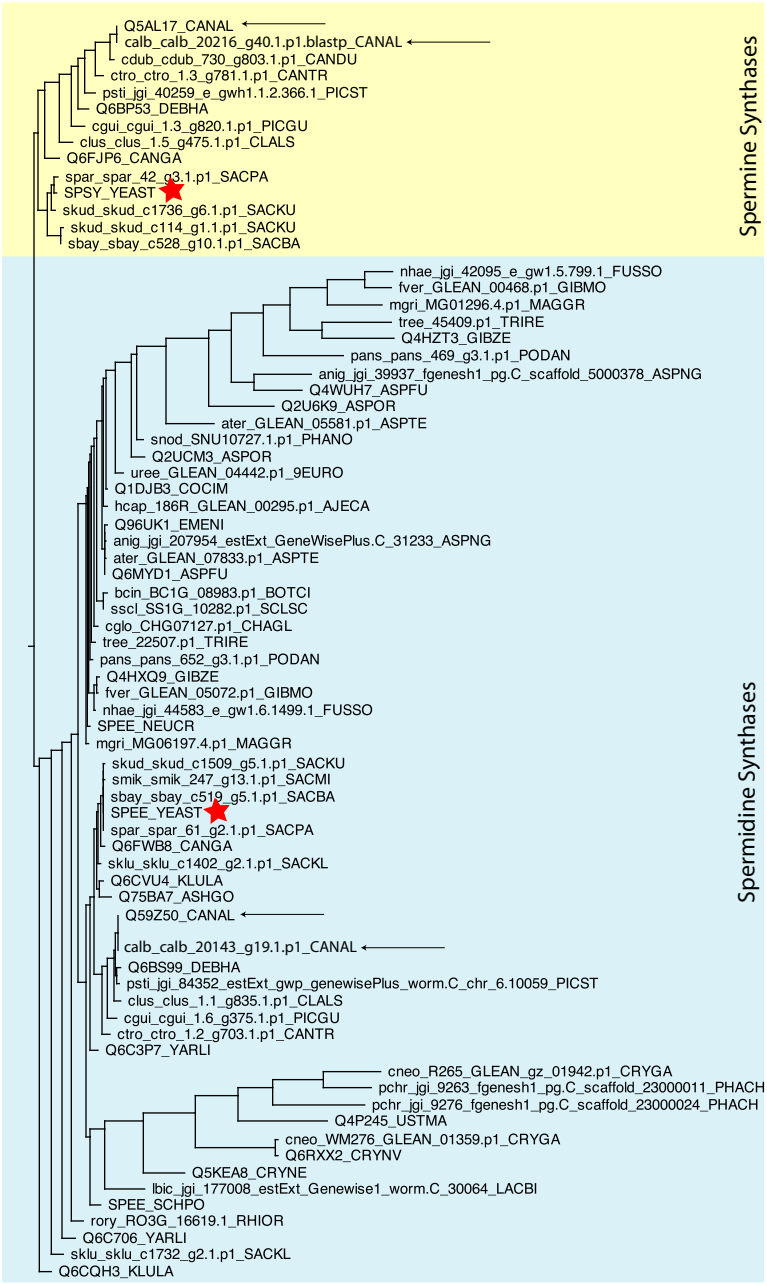

Figure 4.6: Spermine/spermidine synthase tree, with SIFTER's predictions overlaid. The red stars indicate the locations of the two proteins with experimental evidence; the arrows point to the proteins from species discussed in the text.

lowed. Three possible reasons that the spermine synthase clade may be much smaller than the spermidine synthase clade is that either the spermine synthases were lost at a much higher rate, the spermine synthases underwent a functional mutation, or the spermidine synthases were allowed to duplicate more easily. Since the small number of spermine synthases appear to be almost entirely self-contained within a clade of the species tree, it is possible that there were actually only a few gene losses in this clade following the original gene duplication and subsequent functional mutation from a spermidine synthase protein to a spermine synthase protein.

This may be an example of gene *duplicability* being positively correlated or uncorrelated with essentiality in this particular protein family, in contrast to some recent studies that see a negative correlation between gene essentiality and duplicability across a wide range of protein families (e.g., [He and Zhang, 2006]). Another interesting follow-up experiment would be to test whether the subset of faster-evolving spermidine synthase proteins, as indicated by longer branch lengths, are non-essential in their associated species or have a distinct molecular function, both of which may be indicated by their apparent faster rate of mutation.

4.3.4.4 Glycotransferase-34

Phylogenomic analysis of the glycotransferase 34 protein family gives some insight into recent duplication events in this tree. Glycotransferase 34 proteins are involved in a key post-translational modification required for protein secretion. This family is interesting to study for two reasons. First, while different functional subtypes have been identified, few members have experimentally determined function. Second, the fungal organism *Schizosaccharomyces pombe* (which has a comparatively small genome size) has seven proteins in its genome with a glycotransferase domain, whereas most other fungal organisms have many fewer proteins with this domain. In addition to the SIFTER analysis, we performed a BLAST analysis to briefly compare the two

methods and attempt to confirm SIFTER's predictions.

Inspection of annotations retrieved from the GOA database shows that four proteins in this tree contain experimentally supported functions (MNN10_YEAST, MNN11_YEAST, GMA12_SCHPO, and GMH3_SCHPO); these annotations were confirmed through examination of associated literature references. We ran SIFTER on the resulting reconciled tree using the set of experimental annotations; the reconciled phylogeny and the SIFTER predictions are all illustrated in Figure 4.7. In this figure we see that the *S. pombe* sequences are denoted with red stars: our phylogenomic analysis gives us the additional information that, of the seven *S. pombe* glycotransferase 34 proteins, five are predicted to be in the smaller clade containing alpha-1,2-galactosyltransferases, and are the result of three recent duplication events (inparalogs).

For a second functional analysis, we applied BLAST to the collection of sequences generated from the HMM search to perform pairwise annotation transfer for each of the members of this family. On the set of 48 proteins in the family that are contained in the UniProt database, BLAST failed to annotate three UniProt proteins, Q4HU91_GIBZE, Q6C434_YARLI, and Q5BGV8_EMENI (no mannosyltransferase or galactosyltransferase annotations for any proteins in the NR data set with an E-value less than or equal to 0.01). The annotations derived from BLAST (including the three unannotated proteins) differed from the SIFTER predictions for eleven proteins out of the 48 UniProt proteins, or 23%, with one obvious clustering on the phylogeny of these inconsistent predictions (the clade containing the UniProt proteins Q0U0N0_PHANO, Q4HU91_GIBZE, and Q6C434_YARLI, all three of which are either not predicted by BLAST or have inconsistent predictions between BLAST and SIFTER).

The phylogenomic analysis shed light on the important role of galactotransferase proteins in *S. pombe*. One way in which *S. pombe* differs from *S. cerevisiae* is in that its glycoproteins contain both D-mannose and D-galactose, whereas *S. cerevisiae*

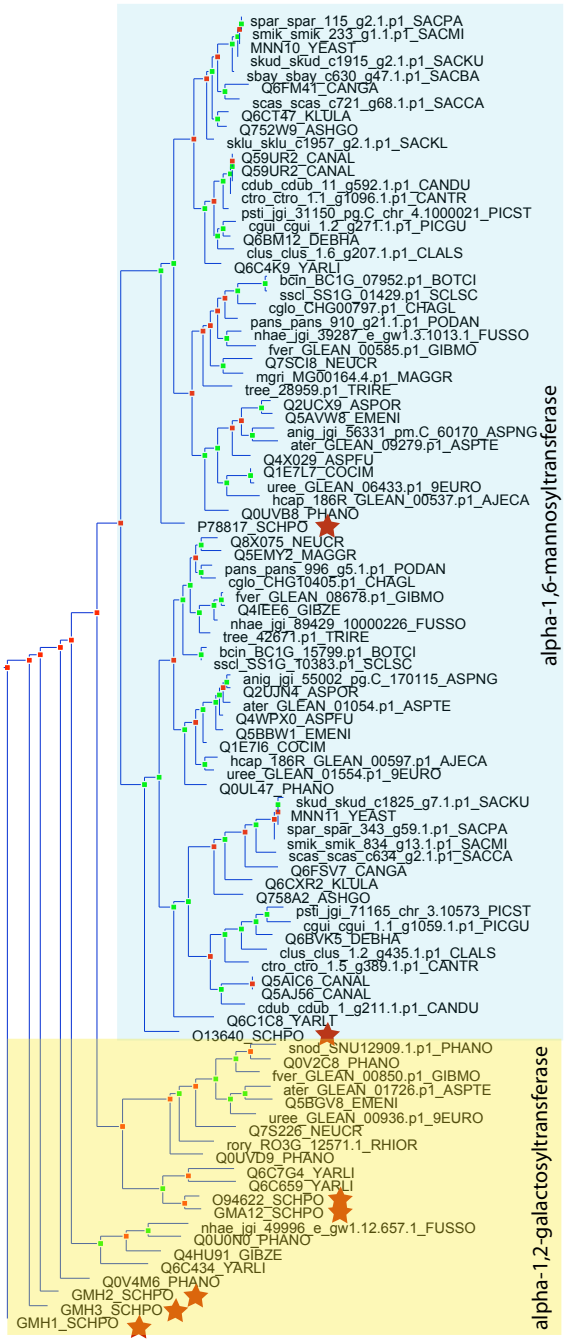

Figure 4.7: Glycotransferase-34 tree, with SIFTER's predictions overlaid. The red stars indicate the *S. pombe* proteins in this family.

glycoproteins contain only D-mannose. Thus, in order for protein glycosylation to occur in *S. pombe*, protein galactosylation must be performed at various locations in the cell. To facilitate this task, it appears that *S. pombe* duplicated additional galactotransferases that localize to different areas of the cell and have distinct but overlapping galactosylation functions [Yoko-o *et al.*, 1998].

4.4 Conclusion

The SIFTER method was applied to perform phylogenomic analyses on all of the protein families found in 46 fully-sequenced fungal genomes, resulting in 2389 protein families run on SIFTER. We described the overall information gained from the SIFTER predictions for these genomes. We present four specific examples of phylogenomic analyses of individual protein families, illustrating the broad power of phylogenomic analyses and phylogenomic functional prediction on a genomic scale.

Chapter 5

Thesis Conclusion

The goal of this thesis is to introduce a statistical method for protein molecular function prediction, including motivating the model itself, presenting applications of this method including a large-scale comparative genomic application, and describing an extension to the basic model that enables active learning. In the Introduction, we review the current literature on how and why molecular function evolves in proteins. This review describes and motivates the phylogenomic methodology upon which we develop the actual statistical method in Chapter 2.

In Chapter 2, we describe how we generate the input data from a database of functional annotations and protein families, and we discuss the fundamentals of the model and the semantics of the model parameters. We show how results from this method, statistical inference of function through evolutionary relationships (SIFTER), compare with the most frequently used method for protein function annotation, BLAST, and two other function annotation methods, GOtcha and Orthostrapper. We show the result of applying SIFTER to three protein families, the AMP/adenosine deaminase family, the sulfotransferase family, and the Nudix family, of which the latter two are functionally diverse. For each of these three protein families, we assembled

a gold-standard set of functional annotations derived from experimental methods from a manual literature search. We present data to validate the truncation approximation for computing the posterior probability predictions on two families, the AMP/adenosine deaminase family and the sulfotransferase family.

In Chapter 3, we motivate the development of an experimental design method to extend SIFTER in terms of the actual cost of performing manual experiments. We describe our simple, information theoretic metric for sequentially selecting a protein to characterize that will maximally reduce prediction uncertainty for the unannotated proteins in the family. We show the results of applying the experimental design to three families with gold-standard molecular function annotation data sets, including the AMP/adenosine deaminase family, the sulfotransferase family, and the aminotransferase family. For the aminotransferase family, presented for the first time here, we also include basic SIFTER results, and we compare our sequential selection of proteins with the selection of proteins via an alterative method, and it appears that the two method produce slightly anti-correlated results. Based on good results for the AMP/adenosine deaminase family and the sulfotransferase family, and inconclusive results for the aminotransferase family, we find that this type of active learning appears to work well in cases where the general phylogenomic technique works well. Specifically, the experimental design appears to work well for single- and multifunction protein families of increasing functional diversity, but has similar difficulty as the general phylogenomic method with sparse annotation data when there is extensive parallel evolution in the family, as in the aminotransferases.

Chapter 4 describes a large-scale application of the SIFTER method to 46 fully-sequenced fungal genomes. The main purpose of this exercise was to determine whether SIFTER could be applied to genome-wide data sets. SIFTER was applied to this large data set and we describe the overall information gained from the brief comparative genomic analysis of these genomes. We further present four examples of

phylogenomic analyses on specific functional domains within this data set, illustrating the broad power of phylogenomic analyses and also indicating the value of functional prediction using phylogenomic analyses on a genome-wide scale.

In conclusion, this simple statistical model of the phylogenomic methodology is effective and robust for molecular function prediction on both individual protein families and genome-wide data sets, and can be used to guide selection of proteins to functionally characterize.

Bibliography

[Allai-Hassani *et al.*, 2007] A Allai-Hassani, P W Pan, L Dombrovski, R Najmanovich, W Tempel, A Dong, P Loppnau, F Martin, J Thonton, A M Edwards, A Bochkarev, A N Plotnikov, M Vedadi, and C H Arrowsmith. Structural and chemical profiling of human cytosolic sulfotransferases. *PLoS Biology*, 5(5):e97, April 2007.

[Alm *et al.*, 2006] E Alm, K Huang, and A Arkin. The evolution of two-component systems in bacterial reveals different strategies for niche adaptation. *PLoS Computational Biology*, 2(11):e143, November 2006.

[Altschul *et al.*, 1990] S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[Altschul *et al.*, 1997] S F Altschul, T L Madden, A A Schaffer, J Zhang, Z Zhang, W Miller, and D J Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25:3389–3402, 1997.

[Amoutzias *et al.*, 2004] G D Amoutzias, D L Robertson, S G Oliver, and E Bornberg-Bauer. Convergent evolution of gene networks by single-gene duplications in higher eukaryotes. *European Molecular Biology Organization Reports*, 5(3):274–279, Feb 2004.

[Andreeva *et al.*, 2004] A Andreeva, D Howorth, S E Brenner, T J P Hubbard, C Chothia, and A G Murzin. Scop database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Research*, 32:D226–D229, 2004.

[Anisimova and Gascuel, 2006] M Anisimova and O Gascuel. Approximate likelihood-ratio test for branches, a fast, accurate, and powerful alternative. *Systematic Biology*, 55:539–552, 2006.

[Apweiler *et al.*, 2004] R Apweiler, A Bairoch, C H Wu, W C Barker, B Boeckmann, S Ferro, E Gasteiger, H Huang, R Lopez, M Magrane, M J Martin, D A Natale,

C O'Donovan, N Redaschi, and L S Yeh. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 32(D):D115–D119, January 2004.

[Arvestad *et al.*, 2003] L Arvestad, A C Berglund, J Lagergren, and B Sennblad. Bayesian gene/species reconciliation and orthology analysis using MCMC. *Bioinformatics*, 19:i7–i15, 2003.

[Ashburner *et al.*, 2000] M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, and J T Eppig. Gene ontology: Tool for the unification of biology. the gene ontology consortium. *Nature Genetics*, 25:25–29, 2000.

[Ashby and Houmard, 2006] M K Ashby and J Houmard. Cyanobacterial two-component proteins: Structure, diversity, distribution, and evolution. *Microbiology and Molecular Biology Review*, 70(2):472–509, June 2006.

[Atteson, 1997] K Atteson. The performance of the NJ method of phylogeny reconstruction. In S. Roberts and A. Rzhetsky, editors, *Mathematical hierarchies and biology*, page 133, Providence, RI, 1997.

[Babushok *et al.*, 2007] D V Babushok, E M Ostertag, and H H Kazazian. Current topics in genome evolution: Molecular mechanisms of new gene formation. *Cellular and Molecular Life Sciences*, 64(5):542–554, March 2007.

[Baker and Sali, 2001] D Baker and A Sali. Protein structure prediction and structural genomics. *Science*, 294(5540):93–96, Oct 2001.

[Bateman *et al.*, 2002] A Bateman, E Birney, L Cerruti, R Durbin, L Etwiller, S R Eddy, S Griffiths-Jones, K L Howe, M Marshall, and E L Sonnhammer. The Pfam protein families database. *Nucleic Acids Research*, 30:276–280, 2002.

[Berglund-Sonnhammer *et al.*, 2006] A-C Berglund-Sonnhammer, P Steffansson, M J Betts, and D A Liberles. Optimal gene trees from sequences and species trees using a soft interpretation of parsimony. *Journal of Molecular Evolution*, 63(2):240–250, August 2006.

[Bergsten, 2005] J Bergsten. A review of long-branch attraction. *Cladistics*, 21:163–193, Feb 2005.

[Bertsekas, 1999] D P Bertsekas. *Nonlinear Programming*. Athena Scientific, New Hampshire, USA, 2nd edition, 1999.

[Bessman *et al.*, 1996] M J Bessman, D N Frick, and S F O'Handley. The MutT proteins or "nudix" hydrolases, a family of versatile, widely distributed, "housecleaning" enzymes. *J Biol Chem*, 271(41):25059–25062, Oct 1996.

[Birney *et al.*, 2004] E Birney, M Clamp, and R Durbin. GeneWise and genomewise. *Genome Research*, 14(5):988–995, 2004.

[Blount and Chatterjee, 1998] B Blount and S Chatterjee. An evaluation of java for numerical computing. In *ISCOPE*, pages 35–46, 1998.

[Bongioanni *et al.*, 1996] P Bongioanni, C Mondino, B Boccardi, M Borgna, and M Castagna. Monoamine oxidase molecular activity in platelets of parkinsonian and demented patients. *Neurodegeneration*, 5(4):339–350, December 1996.

[Bork and Koonin, 1998] P Bork and E V Koonin. Predicting functions from protein sequences – where are the bottlenecks? *Nature Genetics*, 18:313–318, 1998.

[Bowman and Bertozzi, 1999] K G Bowman and C R Bertozzi. Carbohydrate sulfotransferases: Mediators of extracellular communication. *Chemistry & Biology*, 6:9–22, January 1999.

[Brenner, 1999] S E Brenner. Errors in genome annotation. *Trends Genet*, 15:132–133, 1999.

[Brown and Sjolander, 2006] D Brown and K Sjolander. Functional classification using phylogenomic inference. *PLoS Computational Biology*, 2(6):e77, June 2006.

[Camon *et al.*, 2004] E Camon, M Magrane, D Barrell, V Lee, E Dimmer, J Maslen, D Binns, N Harte, R Lopez, and R Apweiler. The gene ontology annotation (GOA) database: sharing knowledge in uniprot with gene ontology. *Nucleic Acids Research*, 32:262–266, 2004.

[Cassola *et al.*, 2004] A Cassola, M Parrot, S Silberstein, B B Magee, S Passeron, L Giasson, and M L Canatore. *Candida albicans* lacking the gene encoding the regulatory subunit of protein kinase a displays a defect in hyphal formation and an altered localization of the catalytic subunit. *European Molecular Biology Organization Journal*, 3:190–199, 2004.

[Chai *et al.*, 2002] W Chai, J G Beeson, and A Lawson. The structural motif in chondroitin sulfate for adhesion of *Plasmodium falciparum*-infected erythrocytes comprises disaccharide units of 4-o-sulfated and non-sulfated n-acetylgalactosamine linked to glucuronic acid. *Journal of Biological Chemistry*, 277(25):22438–22446, June 2002.

[Charlab *et al.*, 2000] R Charlab, E D Rowton, and J M C Ribeiro. The salivary adenosine deaminase from the sand fly. *Experimental Parasitology*, 95:45–53, 2000.

[Chattopadhyay *et al.*, 2002] M K Chattopadhyay, C W Tabor, and H Tabor. Absolute requirement of spermidine for growth and cell cycle progression of fission yeast *Schizosaccharomyces pombe*. *Proceedings of the National Academy of Science U S A*, 99:10330–10334, 2002.

[Chen and Rost, 2002] C P Chen and B Rost. Long membrane helices and short loop. *Protein Science*, 11(12):2766–2773, Dec 2002.

[Chen *et al.*, 2007] F Chen, A J Mackey, J K Vermunt, and D S Roos. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS ONE*, 2(4):e383, Apr 2007.

[Choi and Lahn, 2003] S S Choi and B T Lahn. Adaptive evolution of *MRG*, a neuron-specific gene family implicated in nociception. *Genome Research*, 13:2252–2259, 2003.

[Conrad and Antonarakis, 2007] B Conrad and S E Antonarakis. Gene duplication: A drive for phenotypic diversity and cause of human disease. *Annual Review of Genomics and Human Genetics*, 8(2):2.1–2.19, March 2007.

[Crosby *et al.*, 2007] M A Crosby, J L Goodman, P Strelets, V B Zhang, W M Gelbart, and the FlyBase Consortium. Flybase: genomes by the dozen. *Nucleic Acids Research*, 35:D486–D491, 2007.

[Dayhoff *et al.*, 1978] M O Dayhoff, R M Schwartz, and Orcutt B C. A model of evolutionary change in proteins. In MO Dayhoff, editor, *Atlas of Protein Sequence and Structure*, pages 345–352, 1978.

[Decottignies *et al.*, 2003] A Decottignies, I Sanchez-Perez, and P Nurse. *Schizosaccharomyces pombe* essential genes: A pilot study. *Genome Research*, 13:399–406, 2003.

[Delsuc *et al.*, 2005] F Delsuc, H Brinkmann, and H Philippe. Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics*, 6(5):361–375, May 2005.

[DeLuca *et al.*, 2006] T F DeLuca, I H Wu, J Pu, T Monaghan, L Peshkin, S Singh, and D P Dennis. Roundup: a multi-genome repository of orthologs and evolutionary distances. *Bioinformatics*, 22(16):2044–2046, Jun 2006.

[Devos and Valencia, 2000] D Devos and A Valencia. Practical limits of function prediction. *Proteins*, 41:98–107, 2000.

[Do *et al.*, 2005] C B Do, M S P Mahabhashyam, M Brudno, and S Batzoglou. Probcons: probabilistic consistency-based multiple sequence alignment. *Genome Research*, 15:330–340, 2005.

[Eddy, 1998] S R Eddy. Profile hidden markov models. *Bioinformatics*, 14:755–763, 1998.

[Edgar and Batzoglou, 2006] R C Edgar and S Batzoglou. Multiple sequence alignment. *Current Opinion in Structural Biology*, 16:368–373, May 2006.

[Edgar and Sjolander, 2003] R C Edgar and K Sjolander. SATCHMO: sequence alignment and tree construction using hidden markov models. *Bioinformatics*, 19:1404–1411, 2003.

[Edgar, 2004] R C Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32:1792–1797, 2004.

[Eisen and Hanawalt, 1999] J A Eisen and P C Hanawalt. A phylogenomics study of DNA repair genes, proteins, and processes. *Mutation Research*, 3:171–213, 1999.

[Eisen, 1998] J A Eisen. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Research*, 8:163–167, 1998.

[Engelhardt *et al.*, 2005] B E Engelhardt, M I Jordan, K E Muratore, and S E Brenner. Protein molecular function prediction by bayesian phylogenomics. *PLoS Computational Biology*, 1:e45, 2005.

[Engelhardt *et al.*, 2006] B E Engelhardt, M I Jordan, and S E Brenner. A graphical model for predicting protein molecular function. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, 2006.

[Engelhardt *et al.*, submitted] B E Engelhardt, M I Jordan, and S E Brenner. Scaling SIFTER to annotate large, functionally diverse protein families. 1, submitted.

[Farris, 1982] J S Farris. Simplicity and informativeness in systematics and phylogeny. *Systematic Zoology*, 31(4):413–444, 1982.

[Felsenstein, 1978] J Felsenstein. Cases in which parsimony or compatability methods will be positively misleading. *Systematic Zoology*, 27:401–410, 1978.

[Felsenstein, 1985] J Felsenstein. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39:783–791, 1985.

[Felsenstein, 1989] J Felsenstein. Phylip – phylogeny inference package (version 32). *Cladsistics*, 5:164–166, 1989.

[Felsenstein, 2003] J Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Inc., 2003.

[Ferreira *et al.*, 2005] M E Ferreira, A L Colombo, I Paulsen, Q Ren, J Wortman, J Huang, M H Goldman, and G H Goldman. The ergosterol biosynthesis pathway, transporter genes, and azole resistance in *Aspergillus fumigatus*. *Medical Mycology*, 43(S1):S313–S319, 2005.

[Feyzi *et al.*, 1998] E Feyzi, T Saldeen, E Larsson, U Lindahl, and M Salmivirta. Age-dependent modulation of heparan sulfate structure and function. *Journal of Biological Chemistry*, 273(22):13395–13398, May 1998.

[Fisher, 1930] R A Fisher. *The Genetical Theory of Natural Selection*. The Clarendon Press, 1930.

[Fitzpatrick *et al.*, 2006] D A Fitzpatrick, M E Logue, J E Stajich, and G Butler. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC Evolutionary Biology*, 6:99–114, 2006.

[Force *et al.*, 1999] A Force, M Lynch, F B Pickett, A Amores, Y Yan, and J Postelthwait. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151:1531–1545, April 1999.

[Fraser *et al.*, 2002] H B Fraser, A E Hirsh, L M Steinmetz, C Scharfe, and M W Feldman. Evolutionary rate in the protein interaction network. *Science*, 296:750–752, 2002.

[Freund *et al.*, 1997] Y Freund, H S Seung, E Shamir, and N Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, 1997.

[Galperin and Koonin, 1998] M Y Galperin and E V Koonin. Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement, and operon disruption. *In silico Biology*, 1:55–67, 1998.

[Galperin *et al.*, 2006] M Y Galperin, O V Moroz, K S Wilson, and A G Murzin. House cleaning, a part of good housekeeping. *Mol Microbiol*, 59:5–19, 2006.

[Garcia-Hernandez *et al.*, 2002] M Garcia-Hernandez, T Z Berardini, G Chen, D Crist, A Doyle, E Huala, E Knee, M Lambrecht, N Miller, L A Mueller,

S Mundodi, L Reiser, S Y Rhee, R Scholl, J Tacklind, D C Weems, Y Wu, I Xu, D Yoo, J Yoon, and P Zhang. Tair: a resource for integrated arabidopsis data. *Functional and Integrative Genomics*, 2(6):239, 2002.

[Geer *et al.*, 2002] L Y Geer, M Domrachev, D J Lipman, and S H Bryant. Cdart: Protein homology by domain architecture. *Genome Research*, 12:1619–1623, 2002.

[Gelman *et al.*, 2003] A Gelman, J B Carlin, H S Stern, and D B Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, second edition, 2003.

[Gessner, 1997] M O Gessner. Fungal biomass, production and sporulation associated with particulate organic matter in streams. *Limnetica*, 13:33–44, 1997.

[Gibbs *et al.*, 2006] T T Gibbs, S J Russek, and D H Farb. Sulfated steroids as endogenous neuromodulators. *Pharmacology, Biochemistry and Behavior*, 84(4):555–567, Oct 2006.

[Glasner *et al.*, 2007] M E Glasner, J A Gerlt, and P C Babbitt. Mechanisms of protein evolution and their application to protein engineering. *Advances in Enzymology and Related Areas of Molecular Biology*, 75:193–239, 2007.

[Goodman *et al.*, 1979] M Goodman, J Cselusniak, GW Moore, AE Romero-Herrera, and G Matsuda. Fitting the gene lineage into its species lineage: a parsimony strategy illustrated by cladograms constructed from globin sequences. *Systematic Zoology*, 28:132–168, 1979.

[Gu, 2003] X Gu. Evolution of duplicate genes versus genetic robustness against null mutations. *Trends in Genetics*, 19:354–356, 2003.

[Guindon and Gascuel, 2003] S Guindon and O Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52(5):696–704, 2003.

[Hahn *et al.*, 2005] M W Hahn, T De Bie, J E Stajich, C Nguyen, and N Cristianini. Estimating the tempo and mode of family evolution from comparative genomic data. *Genome Research*, 15:1153–1160, 2005.

[Hall, 2005] B G Hall. Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences. *Molecular Biology and Evolution*, 22(3):792–802, 2005.

[Hamasaki-Katagiri *et al.*, 1997] N Hamasaki-Katagiri, C W Tabor, and H Tabor. Spermidine biosynthesis in *Saccharomyces cerevisae*: polyamine requirement of a null mutant of the SPE3 gene (spermidine synthase). *Gene*, 187(1):35–43, 1997.

[Hamasaki-Katagiri *et al.*, 1998] N Hamasaki-Katagiri, Y Katagiri, C W Tabor, and H Tabor. Spermine is not essential for growth of *Saccharomyces cerevisiae*: Identification of the SPE4 gene (spermine synthase) and characterization of a SPE4 deletion mutant. *Gene*, 210(2):195–201, 1998.

[Hanks *et al.*, 1988] S K Hanks, A M Quinn, and T Hunter. The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. *Science*, 241(4861):42–52, 1988.

[Hawksworth, 2004] D L Hawksworth. Fungal diversity and its implication for genetic resource collections. *Studies in Mycology*, 50:9–18, 2004.

[He and Zhang, 2006] X He and J Zhang. Higher duplicability of less important genes in yeast genomes. *Molecular Biology and Evolution*, 23(1):144–151, January 2006.

[Henikoff and Henikoff, 1992] S Henikoff and J G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Science U S A*, 89(22):10915–10919, November 1992.

[Hirschhorn and Ellenbogen, 1986] R Hirschhorn and A Ellenbogen. Genetic heterogeneity in adenosine deaminase (ADA) deficiency: five different mutations in five new patients with partial ADA deficiency. *American Journal of Human Genetics*, 38:13–25, 1986.

[Hirsh and Fraser, 2001] A E Hirsh and H B Fraser. Protein dispensibility and rate of evolution. *Nature*, 411:1046–1049, Jun 2001.

[Hollich *et al.*, 2005] V Hollich, L Milchert, L Arvestad, and E L L Sonnhammer. Assessment of protein distance measures and tree-building methods for phylogenetic tree reconstruction. *Molecular Biology and Evolution*, 22:2257–2264, 2005.

[Hubbard *et al.*, 2006] T J Hubbard, B L Aken, K Beal, B Ballester, M Caccamo, Y Chen, L Clarke, G Coates, F Cunningham, T Cutts, T Down, S C Dyer, S Fitzgerald, J Fernandez-Banet, S Graf, S Haider, M Hammond, J Herrero, R Holland, K Howe, K Howe, N Johnson, A Kahari, D Keefe, F Kokocinski, E Kulesha, D Lawson, I Longden, C Melsopp, K Megy, P Meidl, B Ouverdin, A Parker, A Prlic, S Rice, D Rios, M Schuster, I Sealy, J Severin, G Slater, D Smedley, G Spudich, S Trevanion, A Vilella, J Vogel, S White, M Wood, T Cox, V Curwen, R Durbin, X M Fernandez-Suarez, P Flicek, A Kasprzyk, G Proctor, S Searle, J Smith, A Ureta-Vidal, and E Birney. Ensembl 2007. *Nucleic Acids Research*, page epub, Dec 2006.

[Huelsenbeck and Ronquist, 2001] J P Huelsenbeck and F Ronquist. MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17:754–755, 2001.

[Huelsenbeck, 1995] J P Huelsenbeck. Performance of phylogenetic methods in simulation. *Systems Biology*, 44(1):17–48, Mar 1995.

[Hurst and Smith, 1999] L D Hurst and N G Smith. Do essential genes evolve slowly? *Current Biology*, 9:747–750, 1999.

[Iyer *et al.*, 2002] L M Iyer, E Koonin, and L Aravind. Extensive domain shuffling in transcription regulators of DNA viruses and implications for the origin of fungal APSES transcription factors. *Genome Biology*, 3:12.1 – 12.11, 2002.

[James and et al, 2006] T Y James and et al. Reconstructing the early evolution of fungi using a six-gene phylogeny. *BMC Evolutionary Biology*, 443(7113):818–822, 2006.

[Jeong *et al.*, 2001] H Jeong, S P Mason, A L Barabasi, and Z N Oltvai. Lethality and centrality in protein networks. *Nature*, 411:41–42, May 2001.

[Jordan, 2004] M I Jordan. *An Introduction to Probabilistic Graphical Models*. to be published, 2004.

[Katoh *et al.*, 2002] K Katoh, K Misawa, K Kuma, and T Miyata. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Research*, 30:3059–3066, 2002.

[Kellis *et al.*, 2004] M Kellis, B W Birren, and E S Lander. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature*, 428:617–624, 2004.

[Koonin *et al.*, 1996] EV Koonin, RL Tatusov, and KE Rudd. Protein sequence compairson at genome scale. *Methods in Enzymology*, 266:295–322, 1996.

[Koonin *et al.*, 2001] E V Koonin, K S Makarova, and L Aravind. Horizontal gene transfer in prokaryotes: quantification and classification. *Annual Reviews of Microbiology*, 55:709–742, 2001.

[Koonin, 1993] E V Koonin. A highly conserved sequence motif defining the family of MutT-related proteins from eubacteria, eukaryotes, and viruses. *Nucleic Acids Research*, 21:4847, 1993.

[Krishnamurthy *et al.*, 2007] N Krishnamurthy, D Brown, and K Sjolander. Flowerpower: clustering proteins into domain architecture classes for phylogenomic inference of protein function. *BMC Evolutionary Biology*, 7(S1):S12, Feb 2007.

[Krissinel, 2007] E Krissinel. On the relationship between sequence and structural similarities in proteomics. *Bioinformatics*, 23(6):717–723, Jan 2007.

[Krylov *et al.*, 2003] D M Krylov, Y I Wolf, I B Rogozin, and Koonin E V. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Research*, 13:2229–2235, 2003.

[Kuhner and Felsenstein, 1994] M K Kuhner and J Felsenstein. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution*, 11:459–468, 1994.

[Kummerfeld *et al.*, 2004] S K Kummerfeld, C Vogel, M Madera, M Pacold, and S A Teichmann. Evolution of multi-domain proteins by gene fusion and fission. In *12th International Conference on Intelligent Systems in Molecular Biology*, Aug 2004.

[Laplazaa *et al.*, 2006] J M Laplazaa, B R Torres, Y-S Jinc, and T J Jeffries. *Sh ble* and cre adapted for functional genomics and metabolic engineering of *Pichia stipitis*. *Enzyme and Microbial Technology*, 38:741–747, 2006.

[Lee *et al.*, 2002] Y Lee, R Sultana, G Pertea, J Cho, S Karamycheva, J Tsai, B Parvizi, F Cheung, V Antonescu, J White, I Holt, F Liang, and J Quackenbush. Cross-referencing eukaryotic genomes: TIGR orthologous gene alignments (TOGA). *Genome Research*, 12:493–502, 2002.

[Lewis and Catlett, 1994] D D Lewis and J Catlett. Heterogeneous uncertainty sampling for supervised learning. In William W. Cohen and Haym Hirsh, editors, *Proceedings of ICML-94, 11th International Conference on Machine Learning*, pages 148–156, New Brunswick, US, 1994. Morgan Kaufmann Publishers, San Francisco, US.

[Li *et al.*, 2003] L Li, C J Stoeckert, and D S Roos. OrthoMCL: idenfication of ortholog groups for eukaryotic genomes. *Genome Research*, 13(9):2178–2189, 2003.

[Liu and Grigoriev, 2004] M Liu and A Grigoriev. Protein domains correlate strongly with exons in multiple eukaryotic genomes – evidence of exon shuffling? *Trends in Genetics*, 20(9):399–403, September 2004.

[Lopez-Errasquin *et al.*, 2006] E Lopez-Errasquin, M T Golzales-Jaen, C Callejas, and C Vazquez. A novel MFS transporter encoding gene in *Fusarium verticillioides* probably involved in iron-siderophore transport. *Mycological Research*, 110(9):1102–1110, 2006.

[Lynch and Force, 2000] M Lynch and A Force. The probability of duplicate gene preservation by subfunctionalization. *Genetics*, 154:459–473, January 2000.

[MacKay, 1992] D MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, 1992.

[Mackey *et al.*, 2007] A J Mackey, Q Liu, F C Pereira, and D S Roos. GLEAN: Improved eukaryotic gene prediction by statistical consensus of gene evidence. *in preparation*, 2007.

[Madern, 2002] D Madern. Molecular evolution within the l-malate and l-lactate dehydrogenase super-family. *Journal of Molecular Evolution*, 54(6):825–840, 2002.

[Maier *et al.*, 2001] S A Maier, L Podemski, S W Graham, H E McDermid, and J Locke. Characterization of the adenosine deaminase-related growth factor (ADGF) gene family in *Drosophila*. *Gene*, 280:27–36, 2001.

[Maier *et al.*, 2005] S A Maier, J R Galellis, and H E McDermid. Phylogenetic analysis reveals a novel protein family closely related to adenosine deaminase. *Journal of Molecular Evolution*, 61(6):776–794, Dec 2005.

[Mar *et al.*, 2005] J C Mar, T J Harlow, and M A Ragan. Bayesian and maximum likeihood phylogenetic analyses of protein sequence data under relative branch-length differences and model violation. *BMC Evolutionary Biology*, 5:8–27, 2005.

[Marchler-Bauer *et al.*, 2002] A Marchler-Bauer, A R Panchenko, B A Shoemaker, P A Thiessen, L Y Geer, and S H Bryant. CDD: A database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Research*, 30:281–283, 2002.

[Martin *et al.*, 2004] D M A Martin, M Berriman, and G J Barton. Gotcha: a new method for prediction of protein function assessed by the annotation of seven genomes. *BMC Bioinformatics*, 5:178–195, 2004.

[McLennan, 2006] A G McLennan. The nudix hydrolase superfamily. *Cell Mol Life Sci*, 63:123–143, 2006.

[Mildvan *et al.*, 2005] A S Mildvan, Z Xia, H F Azurmendi, V Saraswat, P M Legler, M A Massiah, S B Gabelli, M A Bianchet, L-W Kang, and L M Amzel. Structures and mechanisms of nudix hydrolases. *Arch Biochem Biophys*, 433:129–143, 2005.

[Mougous *et al.*, 2004] J D Mougous, C J Petzold, R H Senaratne, D H Lee, D L Akey, F L Lin, S E Munchel, M R Pratt, L W Riley, J A Leary, J M Berger, and C R Bertozzi. Identification, function and structure of the mycobacterial sulfotransferase that initiates sulfolipid-1 biosynthesis. *Nature Structural & Molecular Biology*, 11:721–729, August 2004.

[Mugridge *et al.*, 2000] N B Mugridge, D A Morrison, T Jakel, A R Heckeroth, A M Tenter, and A M Johnson. Effects of sequence alignment and structural domains of ribosomal DNA on phylogeny reconstruction for the protozoan family *Sarcocystidae*. *Molecular Biology and Evolution*, 17(12):1843–1853, 2000.

[Mulder *et al.*, 2007] N J Mulder, R Apweiler, T K Attwood, A Bairoch, A Bateman, D Binns, P Bork, V Buillard, L Cerutti, R Copley, E Courcelle, U Das, L Daugherty, M Dibley, R Finn, W Fleischmann, J Gough, D Haft, N Hulo, S Hunter, D Kahn, A Kanapin, A Kejariwal, A Labarga, P S Langendijk-Genevaux, D Lonsdale, R Lopez, I Letunic, M Madera, J Maslen, C McAnulla, J McDowall, J Mistry, A Mitchell, A N Nikolskaya, S Orchard, O Orengo, R Petryszak, J D Selengut, C J A Sigrist, P D Thomas, F Valentin, D Wilson, C H Wu, and C Yeats. New developments in the InterPro database. *Nucleic Acids Research*, 35:D224–D228, Jan 2007.

[Muratore *et al.*, 2007] K E Muratore, J R Srouji, M A Chow, and J F Kirsch. Recombinant expression of twelve evolutionary diverse subfamily Iα aminotransferases. *Protein Expression and Purification*, 2007.

[Nembaware *et al.*, 2002] V Nembaware, K Crum, J Kelso, and C Seoighe. Impact of the presence of paralogs on sequence divergence in a set of mouse-human orthologs. *Genome Research*, 12:1370–1376, 2002.

[O'Brien *et al.*, 2005] K P O'Brien, M Remm, and E L L Sonnhammer. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Research*, 33:D476–D480, January 2005.

[Ohama *et al.*, 1993] T Ohama, T Suzuki, M Mori, S Osawa, T Ueda, K Watanabe, and T Nakase. Non-universal decoding of the leucine codon CUG in several *Candida* species. *Nucleic Acids Research*, 21(17):4039–4045, 1993.

[Ohno, 1972] S Ohno. *Evolution by Gene Duplication*. Springer-Verlag, 1972.

[Page, 1998] R D M Page. Genetree: comparing gene and species phylogenies using reconciled trees. *Bioinformatics*, 14:819–820, 1998.

[Pal *et al.*, 2003] C Pal, B Papp, and I D Hurst. Genomic function: rate of evolution and gene dispensibility. *Nature*, 421:496–497, 2003.

[Pasek *et al.*, 2006] S Pasek, J-L Risler, and P Brezellec. Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. *Bioinformatics*, 22(12):1418–1423, 2006.

[Pegg *et al.*, 2006] S C Pegg, S D Brown, S Ojha, J Seffernick, E C Meng, J H Morris, P J Chang, C C Huang, T E Ferrin, and P C Babbitt. Leveraging enzyme structure-function relationships for functional inference and experimental design: the structure-function linkage database. *Biochemistry*, 45(8):2545–2555, Feb 2006.

[Perez-Martin *et al.*, 1999] J Perez-Martin, J A Uria, and A D Johnson. Phenotypic switching in *Candida albicans* is controlled by a SIR2 gene. *European Molecular Biology Organization Journal*, 18:2580–2592, 1999.

[Porat *et al.*, 2005] Z Porat, N Wender, O Erez, and C Kahana. Mechanism of polyamine tollerance in yeast: novel regulators and insights. *Cellular and Molecular Life Sciences*, 62(24):3016–3116, 2005.

[Pronzato, 2000] L Pronzato. Adaptive optimization and d-optimum experimental design. *The Annals of Statistics*, 28(6):1743–1761, 2000.

[Ranatunga *et al.*, 2004] W Ranatunga, E E Hill, J L Mooster, E L Holbrook, U Schulze-Gahmen, W L Xu, M J Bessman, S E Brenner, and S R Holbrook. Structural studies of the nudix hydrolase dr1025 from *Deinococcus radiodurans* and its ligand complexes. *Journal of Molecular Biology*, 339:103–116, 2004.

[Reeck *et al.*, 1987] G R Reeck, C de Haen, D C Teller, R F Doolittle, W M Fitch, R E Dickerson, P Chambon, A D McLachlan, E Margoliash, and T H Jukes. "Homology" in proteins and nucleic acids: a terminology muddle and a way out of it. *Cell*, 50:667, 1987.

[Remm *et al.*, 2001] M Remm, C E Storm, and E L Sonnhammer. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology*, 314(5):1041–1052, December 2001.

[Ribard *et al.*, 2003] C Ribard, M Rochet, B Labedan, B Daignan-Fornier, P Alzari, C Scazzocchio, and N Oestreicher. Sub-families of alpha/beta barrel enzymes: a new adenine deaminase family. *Journal of Molecular Biology*, 334:1117–1131, 2003.

[Rocha and Danchin, 2004] E P Rocha and A Danchin. An analysis of determinants of amino acid substitution rates in bacterial proteins. *Molecular Biology and Evolution*, 21(1):108–116, Jan 2004.

[Roth *et al.*, 2007] C Roth, S Rastogi, L Arvestad, K Dittmar, S Light, D Ekman, and D Liberles. Evolution after gene duplication: Models, mechanisms, sequences, systems, and organisms. *Journal of Experimental Zoology (Molecular Development and Evolution)*, 306B:58–73, 2007.

[Saier, 1996] M H Saier. Phylogenetic approaches to the identification and characterization of protein families and superfamilies. *Microbial Comparative Genomics*, 1(3):129–150, 1996.

[Salamov and Solovyev, 2000] A A Salamov and V V Solovyev. Ab initio gene finding in drosophila genomic dna. *Genome Research*, 10(4):516, 2000.

[Sasaki *et al.*, 1995] J Sasaki, L S Brown, Y S Chon, H Kandori, A Maeda, R Needleman, and J K Lanyi. Conversion of bacteriorhodopsin into a chloride ion pump. *Science*, 269:73–75, Jul 1995.

[Scheibel *et al.*, 1998] T Scheibel, T Weikl, and J Buchner. Two chaperone sites in Hsp90 differing in substrate specificity and atp dependence. *Proceedings of the National Academy of Sciences U S A*, 95:1495–1499, 1998.

[Schmit and Mueller, 2007] J Schmit and G Mueller. An estimate of the lower limit of global fungal diversity. *Biodiversity and Conservation*, 16:99–111, 2007.

[Schwarz *et al.*, 2006] E M Schwarz, I Antoshechkin, C Bastiani, T Bieri, D Blasiar, P Canaran, J Chan, N Chen, W J Chen, P Davis, T J Fiedler, L Girard, T W Harris, E E Kenny, R Kishore, D Lawson, R Lee, H-M Müller, C Nakamura, P Ozersky, A Petcherski, A Rogers, W Spooner, M A Tuli, K Van Auken, D Wang, R Durbin, J Spieth, Stein L D, and P W Sternberg. Wormbase: better software, richer content. *Nucleic Acids Research*, 34:D475–478, 2006.

[Searls, 2003] D B Searls. Pharmacophylogenomics: Genes, evolution and drug targets. *Nature Reviews*, 2(8):613–623, 2003.

[Semba *et al.*, 2006] S Semba, S-Y Han, H R Qin, K A McCorkell, D Iliopoulous, Y Pekarsky, T Druck, F Trapasso, C M Croce, and K Huebner. Biological functions of the mammalian Nit1, the counterpart of the invertebrate NitFhit rosetta stone protein, a possible tumor supressor. *Journal of Biological Chemistry*, 281(38):28244–28253, Sep 2006.

[Seung *et al.*, 1992] H S Seung, M Opper, and H Sompolinsky. Query by committee. In Morgan Kaufmann, editor, *Fifth Workshop on Computational Learning Theory*, pages 287–294, 1992.

[Sheehan *et al.*, 2007] M J Sheehan, L M Kennedy, D E Costich, and T P Brutnell. Subfunctionalization of *PhyB1* and *PhyB2* in the control of seedling and mature plant traits in maize. *The Plant Journal*, 49(2):338–353, January 2007.

[Shimodaira, 2002] H Shimodaira. An approximately unbiased test of phylogenetic tree selection. *Systems Biology*, 51:492–508, 2002.

[Sjölander, 2004] K Sjölander. Phylogenomics inference of protein molecular function: advances and challenges. *Bioinformatics*, 20:170–179, 2004.

[Sriram *et al.*, 2005] G Sriram, J A Martinez, E R B McCabe, and K M Dipple. Single-gene disorders: what role could moonlighting enzymes play? *American Journal of Human Genetics*, 76:911–924, 2005.

[Stajich, 2006] J E Stajich. *A comparative genomic investigation of fungal genome evolution*. PhD thesis, Duke University, 2006.

[Storm and Sonnhammer, 2002] C E Storm and E L Sonnhammer. Automated ortholog inference from phylogenetic trees and calculation of ortholog reliability. *Bioinformatics*, 18:92–99, 2002.

[Straney *et al.*, 2002] D Straney, R Khan, R Tan, and S Bagga. Host recognition by pathogenic fungi through plant flavonoids. *Advances in Experimental Medicine and Biology*, 505:9–22, 2002.

[Swofford, 2001] D Swofford. *PAUP*: Phylogenetic Analysis Using Parsimony (*and other methods)*. Sinauer Associates, 2001.

[Tachihara *et al.*, 2005] K Tachihara, T Uemura, K Kashiwagi, and K Igarashi. Excretion of putrescine and spermidine by the protein encoded by ykl174c (tpo5) in *Saccharomyces cerevisiae*. *The Journal of Biological Chemistry*, 280(13):12637–12642, 2005.

[Tagaya *et al.*, 1997] Y Tagaya, G Kurys, T A Thies, J M Losi, N Asimi, J A Hanover, R N Bamford, and T A Waldmann. Generation of secretable and nonsecretable interleukin 15 isoforms through alternate usage of signal peptides. *Proceedings of the National Academy of Sciences U S A*, 94(26):14444–14449, Dec 1997.

[Tateno *et al.*, 1994] Y Tateno, N Takezaki, and M Nei. Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. *Molecular Biology and Evolution*, 11:261–277, 1994.

[Tatusov *et al.*, 1997] R L Tatusov, E V Koonin, and D J Lipman. A genomic perspective on protein families. *Science*, 278:631–637, 1997.

[Tatusov *et al.*, 2000] R L Tatusov, M Y Galperin, D A Natale, and E V Koonin. The cog database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*, 28:33–36, 2000.

[Thomas *et al.*, 2006] J A Thomas, J J Welch, M Wollfit, and L Bromhan. There is no universal molecular clock for invertebrates, but rate variation does not scale with body size. *Proceedings of the National Academy of Sciences*, 103(19):7366–7371, May 2006.

[Thompson *et al.*, 1999] J D Thompson, F Plewniak, and O Poch. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Research*, 27(13):2682–2690, 1999.

[Thornton and LaSalle, 2000] J A Thornton and R LaSalle. Gene family evolution and homology: genomics meets phylogenetics. *Annual Review of Genomics and Human Genetics*, 1:41–73, 2000.

[Torgerson and Singh, 2004] D G Torgerson and R S Singh. Rapid evolution through gene duplication and subfunctionalization of the testes-specific $\alpha 4$ proteasome subunits in drosophila. *Genetics*, 168(3):1421–1432, November 2004.

[UniprotConsortium, 2007] UniprotConsortium. The universal protein resource (uniprot). *Nucleic Acids Research*, 35:D193–197, Jan 2007.

[van der Heijden *et al.*, 2007] R T J M van der Heijden, Snel B, V van Noort, and M A Huynen. Orthology prediction at scalable resolution by phylogenetic tree analysis. *BMC Bioinformatics*, 8(83–94):83, Mar 2007.

[Wall *et al.*, 2005] D P Wall, A E Hirsh, H B Fraser, J Kumm, G Giaever, M B Eisen, and M W Feldman. Functional genomic analysis of the rates of protein evolution. *Proceedings of the National Academy of Sciences, U S A*, 102:5483–5488, 2005.

[Wilson *et al.*, 1977] A C Wilson, S S Carlson, and T J White. Biochemical evolution. *Annual Reviews in Biochemistry*, 46:573–639, 1977.

[Wong *et al.*, 1992] S H Wong, S H Low, and W Hong. The 17-residue transmembrane domain of b-galactoside a2,6-sialyltransferase is sufficient for golgi retention. *The Journal of Cell Biology*, 117(2):245–258, Apr 1992.

[Wu *et al.*, 1999] G Wu, A Fiser, B ter Kuile, A Sali, and M Muller. Convergent evolution of *Trichomonas vaginalis* lactate dehydrogenase from malate dehydrogenase. *Proceeding of the National Academy of Sciences, U S A*, 96:6285–6290, 1999.

[Yang *et al.*, 2003a] J Yang, Z Gu, and W-H Li. Rate of protein evolution versus fitness effect of gene deletion. *Molecular Biology and Evolution*, 20(5):771–774, 2003.

[Yang *et al.*, 2003b] J Yang, R Lusk, and W-H Li. Organismal complexity, protein complexity, and gene duplicability. *Proceedings of the National Academy of Sciences U S A*, 100(26):15661–16665, Dec 2003.

[Yoko-o *et al.*, 1998] T Yoko-o, S K Roy, and Y Jigami. Differences in in vivo acceptor specificity of two galactosyltransferases, the gmh3+ and gma12+ gene products from *Schizosaccharomyces pombe*. *European Journal of Biochemistry*, 257(3):630–637, Nov 1998.

[Yooseph *et al.*, 2007] S Yooseph, G Sutton, D B Rusch, A L Halpern, S J Williamson, K Remington, J A Eisen, K B Heidelberg, G Manning, W Li, L Jaroszewski, P Cieplak, C S Miller, H Li, S T Mashiyama, M P Joachimiak, C van Belle, J M Chandonia, D A Soergel, Y Zhai, K Natarajan, S Lee, B J Raphael, V Bafna, R Friedman, S E Brenner, A Godzik, D Eisenberg, J E Dixon, S S Taylor, R L Strausberg, M Frazier, and J C Venter. The *Sorcerer II* global ocean sampling expedition: Expanding the universe of protein families. *PLoS Biology*, 5(3):e16, March 2007.

[Zmasek and Eddy, 2001a] C M Zmasek and S R Eddy. A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics*, 17:821–828, 2001.

[Zmasek and Eddy, 2001b] C M Zmasek and S R Eddy. ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics*, 17:383–384, 2001.

[Zmasek and Eddy, 2002] C M Zmasek and S R Eddy. RIO: Analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*, 3:14, 2002.